The Korean Communications in Statistics Vol. 15, No. 1, 2008, pp. 27–41

일반화선형모형에서 선형성의 타당성을 진단하는 그래프†

김지현[1)]

요 약

그림으로 일반화 선형모형의 적합성을 진단하는 방법을 제안한다. 이 그림은 일반화 선형모형에서 연결함수를 설명변수들의 선형결합으로 표현할 수 있다는 가정을 진단할 때 유용하다. 이 그림에서 연결함수와 설명변수들의 관계를 비모수적으로 추정하는 작업이 필요한데, 이를 위해 여러 가능한 기법 중에서 부스팅 기법을 적용하였다. 정규분포와 이항분포 자료로 모의실험을 실시하여 새로이 제안한 진단그림의 효과성을 보였다. 그리고 진단그림의 한계와 기술적 세부사항들을 설명하였다.

주요용어: 부스팅; 비모수적 회귀; 붓스트랩 신뢰구간.

1. 서론

일반화선형모형 (generalized linear models)은 여러 가지 모수적 (parametric) 회귀모형의 통합적인 모형으로 널리 쓰이고 있다. 예측의 정확성을 높이기 위해 비모수적 (nonparametric) 회귀모형에 대한 관심이 높아지고 있으나, 모수적 모형은 적합도 (goodness-of-fit)에 문제만 없다면 간결하고 해석이 쉽다는 장점을 갖는다. 일반화선형모형의 적합도를 통계량을 이용하여 검정하는 방법에 대해서는 많은 연구가 이루어졌으나 (Su와 Wei, 1991; Cheng과 Wu, 1994) 그림으로 진단하는 방법에 대한 연구는 많지 않다. 보통선형회귀 (ordinary linear regression)에서의 잔차를 이용한 진단이나 로지스틱 회귀에 대한 Landwehr 등 (1984)의 연구 등이 있으나 이들 모형을 통합한 일반화선형모형에 적용할 수 있는 진단그림은 제안된 것이 없다. 본 연구에서는 일반화선형모형에 공통적으로 적용할 수 있는 진단그림을 제안한다.

† 이 논문은 2005년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2005-015-C00086).

1) (156-743) 서울시 동작구 상도동, 숭실대학교 정보통계보험수리학과, 교수
E-mail: jxk61@ssu.ac.kr

2절에서 새로운 진단그림의 필요성을 예를 들어 설명하였다. 그리고 그림을 통해 진단하고자 하는 가설이 무엇인가에 대해 설명하고 진단그림을 그리는 방법을 제시하였다. 3절에서 모의실험을 통해 진단그림의 성능을 평가하였으며, 4절에서 진단그림의 한계와 추가적인 몇 가지 사항에 대해 논하였다.

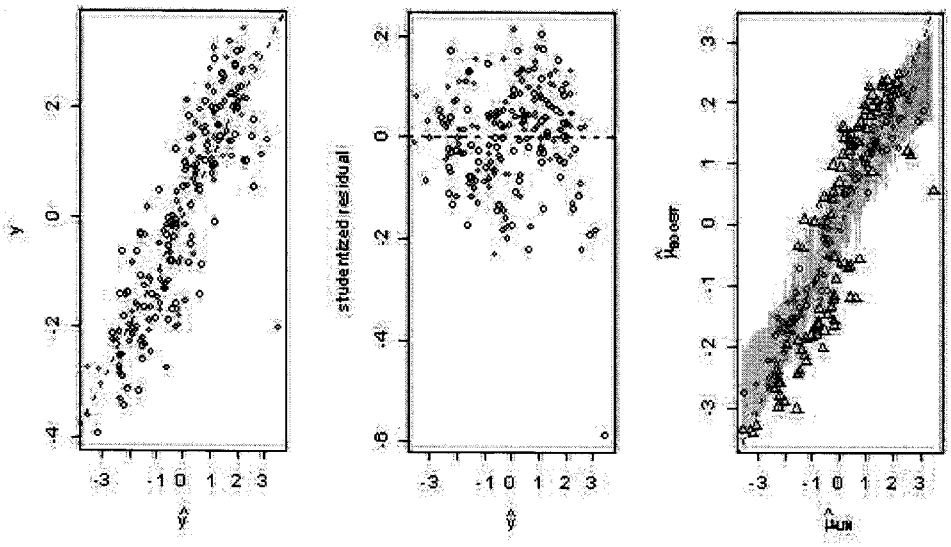
2. 연구의 필요성과 방법의 제안

2.1. 연구의 필요성

통계적 모형의 적합도는 모형의 해석과 이용에 앞서 꼭 살펴보아야 할 부분이다. 본 논문에서는 모형의 적합도 또는 타당성에 대한 진단 중에서 회귀식의 선형성 가정, 좀 더 일반적으로 일반화선형모형에서 연결함수가 설명변수들의 선형결합으로 표현될 수 있다는 가정을 그래프로 진단하는 방법에 대해 연구하였다. 진단방법을 기술하기 전에 그 필요성을 예를 통해 설명하여 본다.

오차항의 정규성을 가정하는 선형회귀모형에서 잔차를 이용한 그림들로 오차항의 등분산성과 독립성, 정규성 등을 진단한다. 회귀식의 선형성을 판단하는 데에 유용한 그림으로 추가변수그림 (added variable plot)과 편잔차그림 (partial residual plot) 등이 있으나, 이 그림들은 관심있는 설명변수를 제외한 나머지 설명변수들의 선형성을 가정하고 있다. 이러한 가정 없이 각 변수들의 선형성 여부를 진단하기 위해서는 축차적으로 변수들을 추가하며 살펴보아야 하는데 이런 번거로운 작업을 하기 전에 전체적으로 그 필요성을 판단하기 위한 그림이 있으면 도움이 될 것이다. 모형의 적합도를 판단하기 위해 그림 2.1의 (a), (b)와 같이 관측값 y와 예측값 $\hat{y}$의 그림 또는 잔차 $y-\hat{y}$와 예측값 $\hat{y}$의 그림을 생각해볼 수 있겠으나, 이들은 오차항을 포함하고 있는 그림이기 때문에 평균함수의 선형성을 판단하는 데에는 적절하지 않다. 왜냐하면 오차항의 분포형태나 등분산성의 만족여부 또는 특이점 (outliers)의 존재 등이 평균함수의 선형성을 판단하는 데에 방해가 될 수 있기 때문이다. 본 연구에서는 그림 2.1의 (c)와 같이 평활을 통해 오차항의 효과를 제거한 그림을 제안한다. 그림 2.1 의 (a)와 (b)를 통해서 모형에 문제가 있다는 것은 짐작해볼 수 있으나, 그것이 오차항의 분산 때문인지 평균함수의 선형성에 문제가 있어서인지는 판단하기가 쉽지 않다. 실제로 이 그림을 위한 자료는 등분산성은 만족되지만 평균함수의 선형성이 만족되지 않는 자료인데, 본 연구에서 제안한 그림 2.1의 (c)는 평균함수의 선형성에 문제가 있다는 것을 분명하게 보여준다. 뒤에서 설명하겠지만 세모모양의 점이 많을수록 모형의 선형성에 문제가 있다는 것을 나타내기 때문이다.

선형회귀모형의 적합도를 그림으로 진단할 때 잔차를 이용하지만 이항자료에



(a)$\hat{y}$ 대 y의 그림 (b) $\hat{y}$ 대 잔차의 그림 (c) 본 연구에서 제안한 그림

그림 2.1: 모형의 적합도 진단을 위한 그림 (설명변수가 10개이고, 등분산성은 성립하나 평균함수의 선형성을 만족하지 않는 자료로서 자료의 크기는 200)

대해서는 자료의 이산성으로 인해 잔차 그림이 적절하지 않다. 이 때 흔히 그룹화(grouping)하여 이산성의 문제를 극복하는데, 설명변수의 수가 많거나 연속형 설명변수가 있을 때는 그룹화가 힘들다. Landwehr 등 (1984)의 국소평균이탈도그림(local mean deviance plot)은 이항자료에 대해 모형의 적합도를 진단하는 그림이지만, 설명변수들로 자료를 군집화한 다음에 찾은 군집을 이용하여 그룹화해야 하므로 번거롭고 군집과정에서의 주관성도 문제가 된다. 본 연구에서 제안한 그림을 이용하면 이항자료에 대해서도 평균함수의 선형성을 진단할 수 있다.

일반화선형모형은 반응변수의 분포를 다양하게 허용하지만 연결함수가 설명변수의 선형결합으로 표현되어야 한다는 제약을 갖고 있다. 모형을 이용하여 해석하거나 추론하기 전에 이 제약의 타당성을 필수적으로 점검해야 한다. 이 때 그림으로 진단할 수 있다면 통계량에 의한 유의성 검정의 한계 (Kim, 2003)를 보완할 수 있어 매우 유용할 것이다.

2.2. 방법의 제안

일반화선형모형은 연속형뿐만 아니라 이산형 반응변수를 아우르는 회귀모형으로서, 반응변수 Y의 분포함수는

$$f(y|\theta,\phi,\boldsymbol{x}) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{2.1}$$

로서 지수족에 속하며, μ를 Y의 평균이라고 할 때,

$$g(\mu) = \boldsymbol{x}'\beta \tag{2.2}$$

즉, 평균에 관한 연결함수 (이하 '평균함수'로 부르기로 함)를 설명변수의 선형결합으로 표현할 수 있다는 가정을 하고 있다. 평균함수의 선형성은 강한 가정으로서 보다 일반적인 모형

$$g(\mu) = F(\boldsymbol{x})$$

과 비교하여 그 타당성을 살펴볼 필요가 있다. 이 때 함수 F는 선형성의 제약이 없는 일반적인 함수이다.

주어진 자료 $\{(y_i,\boldsymbol{x});\ i=1,\ldots,n\}$에 대해 일반화선형모형 (2.1) (2.2)를 가정하였을 때, 분포에 대한 가정 (2.1)과 연결함수 g의 적절성은 만족된다는 전제 하에 평균함수의 선형성 (2.2)를 진단하는 그림 (이하 '선형성진단그림'으로 부르기로 함)을 그리는 방법은 다음과 같다.

(1) 비모수적 회귀(nonparametric regression)로 $\eta = F(\boldsymbol{x})$과 $\mu = g^{-1}(\eta)$를 추정

(2) 일반화선형모형을 가정하고 $\eta = \boldsymbol{x}'\beta$와 $\mu = g^{-1}(\eta)$를 추정

(3) (1)에서 구한 μ의 추정값 $\hat{\mu}_{NP}$와 (2)에서 구한 μ의 추정값 $\hat{\mu}_{LIN}$을 비교하는 그래프를 그림.

평균함수의 선형성이 만족된다면 선형성진단그림에서 자료의 점들이 기울기 1인 직선 근처에 분포할 것으로 기대된다. 하지만 주어진 자료에 대한 선형성진단그림에서 우연성 (randomness) 때문에 선형성의 만족 여부를 판단하기가 쉽지 않다. 즉, 우연성을 제외하면 기울기 1인 직선으로 볼 수 있는 것인지, 아니면 평균함수의 선형성이 만족되지 않아 랜덤하게 보이는 것인지를 판단하는 것이 주관적일 수 있다. 이 때 객관적인 판단을 도와줄 수 있는 장치가 필요한데 다음과 같은 붓스트랩 신뢰구간을 제안한다.

(i) 앞에서 설명한 단계 (2)에서 얻은 $\mu(\boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i'\beta)$, $i = 1, \ldots, n$의 추정값을 $\hat{\mu}_{LIN}(\boldsymbol{x}_i)$ 이라고 할 때 일반화선형모형에서 가정한 임의성분 (random component)의 분포로부터 새로운 반응변수의 값 y_i^*를 생성한다. 예를 들어 정규분포라면 $y_i^* \sim N(\hat{\mu}_{LIN}(\boldsymbol{x}_i), \hat{\sigma}^2)$이고 이항분포라면 $y_i^* \sim b(1, \hat{\mu}_{LIN}(\boldsymbol{x}_i))$이다. 정규분포의 예에서 $\hat{\sigma}^2$은 주어진 자료에 적용한 모형에서 얻어지는 분산의 추정값이다. (등분산성의 제약을 두지 않고 잔차들의 제곱을 평활한 값을 분산의 추정값으로 이용하는 방법도 가능하다.)

(ii) 단계 (i)에서 생성한 자료 $\{(y_i^*, \boldsymbol{x}_i), i = 1, \ldots, n\}$에 비모수회귀를 적용하여 $\hat{\mu}_{NP}^*(\boldsymbol{x}_i)$를 얻는다. 자료 $\{(y_i^*, \boldsymbol{x}_i), i = 1, \ldots, n\}$는 일반화선형모형 $g(\hat{\mu}_{LIN}(\boldsymbol{x}_i)) = \boldsymbol{x}_i'\hat{\beta}$을 따르므로 $\hat{\mu}_{NP}^*(\boldsymbol{x}_i)$는 우연 변동 (random variation)을 제외하고는 $\hat{\mu}_{LIN}(\boldsymbol{x}_i)$와 다르지 않을 것이다.

(iii) 단계 (i)과 (ii)를 B번 반복하여 각 $\boldsymbol{x}_i$에서 B개의 추정값 $\hat{\mu}_{NP}^{*(1)}(\boldsymbol{x}_i), \ldots, \hat{\mu}_{NP}^{*(B)}(\boldsymbol{x}_i)$를 얻는다. 이 값들로부터 선형성 가정이 만족될 때 $\mu(\boldsymbol{x}_i)$에 대한 근사적 신뢰구간을 얻는다. 모든 $\boldsymbol{x}_i$에서 신뢰구간을 구하고 상한은 상한들끼리 하한은 하한들끼리 연결하여 그림에 표시한다.

그림 2.1의 (c)는 평균함수의 선형성이 만족되지 않는 인위적인 자료에 대해 기울기 1인 직선과 붓스트랩 신뢰구간을 함께 표시한 선형성진단그림이다. (이 그림에 쓰인 자료는 다음 절에서 설명할 모의실험 자료인데, 설명변수가 10개이며 오차항이 정규분포를 따르고 등분산성을 만족하지만 선형성은 만족하지 않는다.) 95% 신뢰구간을 벗어나는 점의 비율이 약 0.34로 높고 신뢰구간을 크게 벗어나는 점도 많은 것으로 봐서 평균함수의 선형성에 문제가 있다는 올바른 진단을 할 수 있다.

3. 모의실험

앞 절에서 제안한 선형선진단그림의 효용성을 알아보기 위해 인위적 자료를 생성하여 실험하였다. 설명변수가 1개인 경우와 여러 개인 경우 각각에 대하여, 선형성이 만족될 때와 만족되지 않을 때로 나누어 자료 크기를 달리하며 실험하였다.

선형성진단그림을 그리는 방법 중 단계 (1)에서 비모수적 회귀로 μ를 추정한다고 했는데, 본 연구에서는 나무모형을 약분류기 (weak learners)로 하는 부스팅 (boosting) 기법을 적용하였다. 부스팅은 기계학습 (machine learning) 분야의 연구자들에 의해 분류 문제에서 오분류율을 낮추기 위한 방법으로 제안되었다 (Freund와 Schapire, 1997). 이후 Friedman 등 (2000)는 부스팅 기법을 가능도 (likelihood)를 최대로 하는 함수를 경사법 (gradient method)에 의해 추정하는 방법

으로 간주할 수 있음을 보임으로써 분류문제뿐만 아니라 회귀문제 등 여러 통계 문제에 확장 적용할 수 있는 가능성을 열었다. 이에 주목하여 Ridgeway (1999)는 구체적으로 부스팅을 일반화선형모형과 Cox의 비례위험모형 (proportional hazards model)에 적용하는 방법을 제시하였는데, 본 연구에서는 평균함수 $F(x)$를 추정할 때 Ridgeway의 방법을 이용하였다. R언어로 (R Development Core Team, 2006) 프로그램 하였으며 나무모형 적합을 위해 rpart 팩키지를 썼다 (Therneau와 Atkinson, 2006).

3.1. 설명변수가 1개인 경우의 모의실험

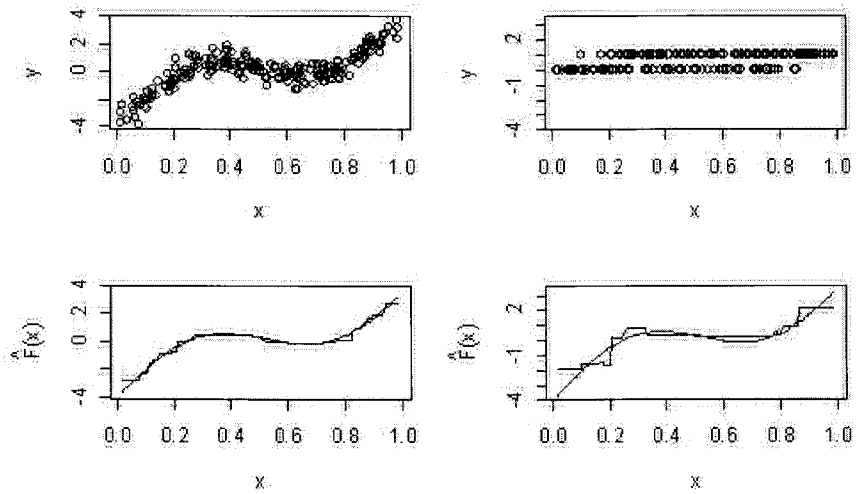
선형성진단그림의 성능을 알아보기 위해 먼저 설명변수가 1개인 경우를 실험하였다. 물론 설명변수가 1개일 때는 굳이 선형성진단그림이 아니더라도 선형성의 타당성 여부를 쉽게 알아볼 수 있지만, 선형성진단그림이 제대로 작동하는지 확인하기 위하여 선형성이 만족되지 않는 경우와 선형성이 만족되는 경우 각각에 대하여 정규분포와 이항분포 자료로 실험하였다.

선형성이 만족되지 않는 경우 선형성진단그림의 성능을 알아보기 위해 다음 식에 따라

$$g(\mu) = F(x) = 8(x - 0.5) + 2\sin(6x), \quad x \sim U(0,1)$$

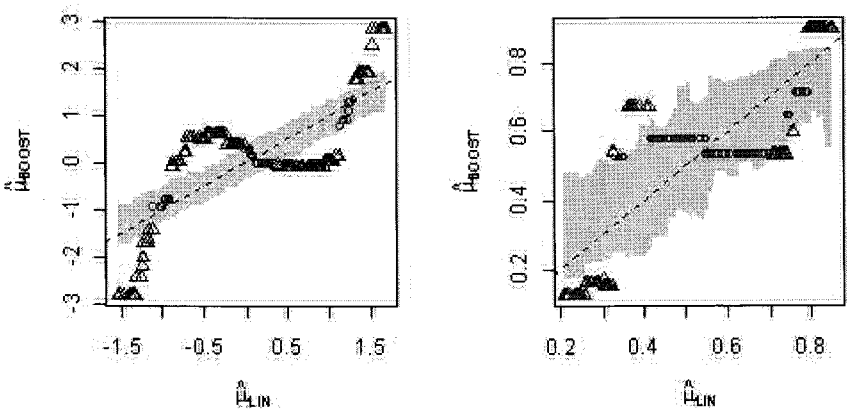
정규분포 $N(\mu, 0.5^2)$, $\mu = F(x)$와 이항분포 $b(1,\mu)$, $\mu = g^{-1}(F(x)) = 1/[1 + \exp(-F(x))]$에서 각각 200개씩의 자료를 생성하여 훈련자료 (training data)를 구성하였다. 그림 3.1에서 자료로부터 y대 x의 그림을, 그리고 부스팅에 의한 추정의 성능을 알아보기 위해 참값 $F(x)$와 부스팅에 의한 추정값 $\hat{F}(x)$의 그림을 그려 보았다. 이항자료의 경우 0 또는 1로 주어지는 자료의 특성상 $F(x)$를 추정하는 것이 정규자료에 비해 어렵다는 것을 알 수 있다. (이 사실은 정규분포의 분산을 0.5^2에서 1과 1.5^2으로 크게 했을 때도 마찬가지였다.)

그림 3.2는 선형성진단그림인데, 이항자료의 경우 신뢰구간이 더 넓게 분포되어 선형성 진단이 정규자료에 비해 어려움을 알 수 있다. 정규자료의 경우 선형성진단그림에서 95% 신뢰구간을 벗어나는 점의 비율이 0.75이었고, 이항자료는 0.445로서 선형성진단그림이 제대로 작동한다는 것을 알 수 있다. 위 실험을 100번 반복하여 얻은 표 3.1의 '비선형자료' 열을 보면 위 실험결과가 우연히 나타난 것이 아니라 지속적인 것임을 알 수 있다. 이 때 약분류기로 쓰는 나무모형의 최대깊이를 1과 2로 달리 하면서 실험하였는데, 두 경우 모두 표본 크기가 커짐에 따라 신뢰구간을 벗어나는 점들의 비율이 높아지는 것을, 즉 검정력이 커지는 것을 알 수 있다.



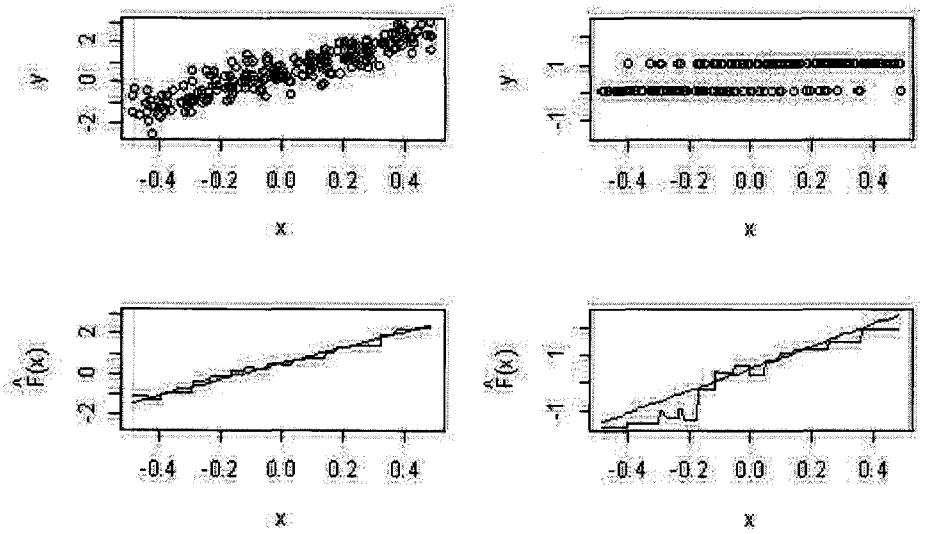
(a) 정규자료와 평균의 추정값 (b) 이항자료와 선형성분의 추정값

그림 3.1: 설명변수가 1개이고 선형성을 만족하지 않는 자료(부드러운 곡선은 참값 $F(x)$)



(a) 정규분포 (b) 이항분포

그림 3.2: 그림 3.1의 자료에 대한 선형성진단그림



(a) 정규자료와 평균의 추정값 (b) 이항자료와 선형성분의 추정값

그림 3.3: 설명변수가 1개이고 선형성을 만족하는 자료 (직선은 참값 $F(x)$)

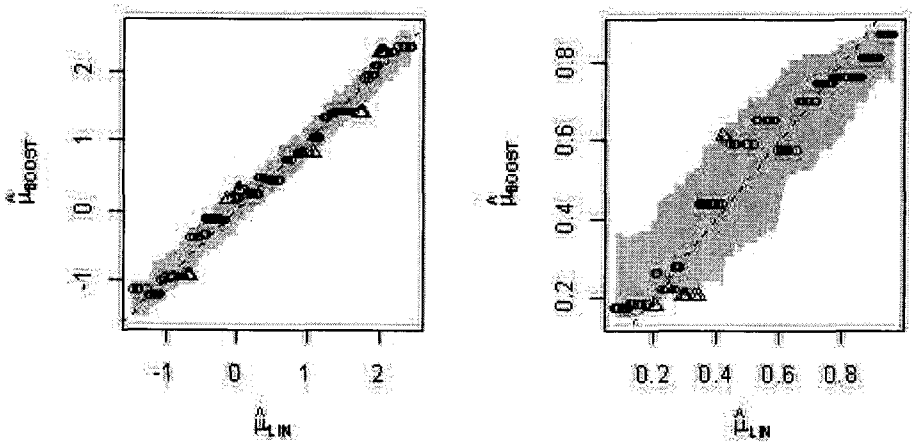


그림 3.4: 그림 3.3의 자료에 대한 선형성진단그림

표 3.1: 설명변수가 하나일 때 95% 신뢰구간을 벗어나는 점의 비율 (괄호 안의 값은 표준오차)

		나무 최대깊이= 1		나무 최대깊이= 2	
분포	표본크기	비선형자료	선형자료	비선형자료	선형자료
정규분포	100	.6213	.0618	.6286	.0585
		(.0057)	(.0054)	(.0071)	(.0053)
	200	.7038	.0621	.7127	.0617
		(.0047)	(.0042)	(.0053)	(.0043)
	400	.7687	.0655	.7805	.0662
		(.0037)	(.0044)	(.0042)	(.0039)
이항분포	100	.2549	.0874	.2428	.1032
		(.0183)	(.0129)	(.0204)	(.0155)
	200	.3860	.0598	.3179	.0774
		(.0179)	(.0092)	(.0181)	(.0123)
	400	.5252	.0922	.4523	.0961
		(.0413)	(.0117)	(.0139)	(.0120)

이번에는 선형성이 만족될 때 선형성진단그림의 성능을 알아보기 위해

$$g(\mu) = F(x) = 0.5 + 4(x - 0.5), \quad x \sim U(0,1) \tag{3.1}$$

를 만족하는 정규자료와 이항자료를 생성하여 그림 3.3와 그림 3.4를 얻었다. 그래프에서 점들이 직선 근처에 분포해 있고 신뢰구간을 벗어나는 점의 비율이 정규분포와 이항분포의 경우 각각 0.095와 0.055로서 작아, 선형성진단그림이 제대로 작동함을 알 수 있다. 같은 실험을 100번 반복하여 얻은 결과를 표 3.1 '선형자료' 열에 수록하였다.

3.2. 설명변수가 여러 개인 경우의 모의실험

설명변수가 많은 모의실험자료를 생성하기 위해 다음과 같은 모형을 고려하였다.

$$F(\boldsymbol{x}) = -3 + 5\sin(\pi x_1 x_2) - 5\left(x_3 - \frac{1}{2}\right)^2 + 0.3x_4 + 0.3x_5 \tag{3.2}$$

처음 4개의 설명변수들은 구간 $(0,1)$에서 균일분포를 따르며, x_5의 분포는 이산형 균일분포로서 가능한 값은 $\{1,2,3\}$이다. 실제 자료에는 반응변수와 관련이 없는 설명변수도 포함될 수 있음을 감안하여, 분석을 위한 자료에는 위 5개의 설명변수에 반응변수와 아무 관련이 없는 5개의 설명변수를 추가하였다. 추가적인 변수들의 분포는 참모형 $F(\boldsymbol{x})$에 있는 5개의 설명변수와 같은 분포를 갖도록 하였다. 반

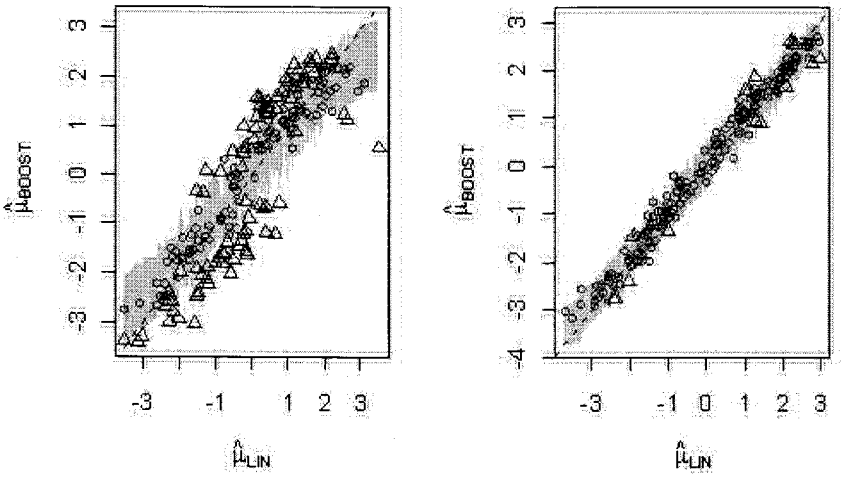
응변수 y는 정규분포 $N(\mu, 0.5^2), \mu = F(\boldsymbol{x})$와 이항분포 $b(1, \mu), \mu = g^{-1}(F(\boldsymbol{x}))$에서 각각 생성하였으며, 훈련표본의 크기는 200이다.

모형 (3.1)로부터 3가지 종류의 자료를 생성하여 실험하였다. 먼저 선형성이 만족되지 않는 경우로서 자료 $\{(y_i, x_{i1}, \ldots, x_{i10}),\ i = 1, \ldots, 200\}$에 대해 실험하였다. $F(\boldsymbol{x})$는 주어진 설명변수 $(x_1, \ldots, x_{10})$에 대해 선형이 아닌데, 그림 3.5 (a)와 그림 3.6 (a)와 같이 선형성이 만족되지 않는다는 진단을 그림을 통해 내릴 수 있다. (95% 신뢰구간을 벗어나는 점의 비율이 각각 0.420, 0.145 이다.) 두 번째 실험자료는 선형성이 만족되는 자료로서 $\{(y_i, z_{i1}, z_{i2}, x_{i4}, \ldots, x_{i10}),\ i = 1, \ldots, 200\}, z_{i1} = \sin(\pi x_{i1}, x_{i2}), z_{i2} = (x_{i3} - 1/2)^2$이다. 이 자료에 대한 선형성진단그림이 그림 3.5 (b)와 그림 3.6 (b)이다. 예상대로 이 자료의 경우 95% 신뢰구간을 벗어나는 점의 비율이 각각 0.08, 0.01로서, 모형의 선형성에 별 문제가 없다고 올바른 진단을 내릴 수 있다.

위 실험결과의 일관성을 보기 위해 같은 실험을 100번 반복하여 표 3.2를 얻었다. 표본의 크기는 설명변수가 1개인 실험과 달리 200, 400, 800으로 정하여 실험하였다. (설명변수가 10개인 모형에서 100개의 자료는 계수 추정을 위해 충분히 큰 자료가 아닌데, 특히 이항분포일 때 $n = 100$이면 수렴하지 않는 경우가 잦았다.) 먼저 신뢰구간을 벗어나는 점의 비율을 살펴보았다. 선형성이 만족될 때의 비율이 그렇지 않을 때에 비해 작음을 알 수 있다. 표본 크기가 커질수록 선형성이 만족될 때와 그렇지 않을 때 신뢰구간을 벗어나는 점의 비율의 차이가 커진다는 사실도 알 수 있다.

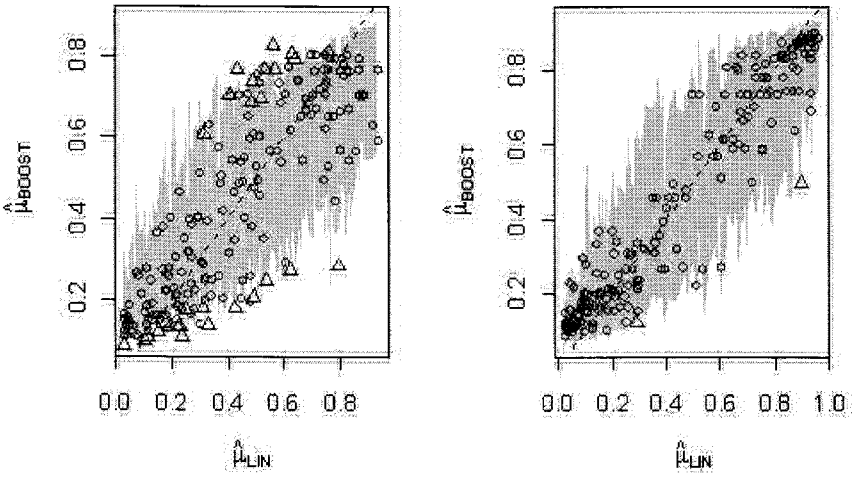
붓스트랩 신뢰구간을 구하기 위해 $B = 100$번 붓스트랩 자료를 생성하였다. Efron과 Tibshirani (1993, p. 162)는 신뢰수준 90% 또는 95%의 붓스트랩 신뢰구간을 구하기 위해서 B를 1000에서 2000 정도의 값으로 할 것을 권장하고 있는데, 본 연구에서는 실행시간의 제약 때문에 100으로 하여 실험하였다. 이항분포이면서 표본의 크기가 200과 400인 경우에 한해서 B를 1000으로 하여 실험해본 결과, 95% 신뢰구간을 벗어나는 점의 비율이 비선형과 선형의 경우 모두 조금씩 낮아짐을 확인할 수 있었다. (비선형인 경우 0.2117, 0.3197은 각각 0.1941, 0.2970으로, 선형인 경우 0.0907, 0.0778은 각각 0.0520, 0.0533으로 낮아짐.)

주어진 자료로부터 모형의 선형성을 그림으로 진단할 때, 단지 신뢰구간을 벗어나는 점의 비율만 보지 말고 신뢰구간을 얼마나 벗어나는가를 같이 보아야 한다. 그리고 한 모형의 타당성을 진단그림으로 판단하기는 어려울 수 있으나, 몇 가지 후보 모형을 비교하여 최적 모형을 찾을 때에는 진단그림이 매우 유용할 것이다. 이 때 $\hat{\mu}_{LIN}$와 $\hat{\mu}_{BOOST}$의 선형상관계수를 구해 진단그림에서 신뢰구간을 벗어나는 정도를 정량적으로 비교하면 도움이 된다.



(a) 선형성 불만족 자료 (b) 선형성 만족 자료

그림 3.5: 설명변수가 여러 개인 정규분포 자료에 대한 선형성진단그림



(a) 선형성 불만족 자료 (b) 선형성 만족 자료

그림 3.6: 설명변수가 여러 개인 이항분포 자료에 대한 선형성진단그림

표 3.2: 여러 개의 설명변수가 있는 두 종류의 자료에 대한 100번의 실험 결과: 표시한 값은 신뢰구간을 벗어나는 점의 비율 (괄호 안의 값은 표준오차)

분포	표본크기	비선형자료	선형자료
정규분포	200	.3769	.0725
		(.0067)	(.0038)
	400	.4707	.0641
		(.0053)	(.0027)
	800	.5519	.0597
		(.0038)	(.0018)
이항분포	200	.2117	.0907
		(.0197)	(.0128)
	400	.3197	.0778
		(.0173)	(.0085)
	800	.3984	.0637
		(.0131)	(.0075)

4. 토의 및 결론

지금까지 일반화선형모형의 선형성을 진단하는 그래프를 제안하고, 진단도구로서의 효용성을 모의자료를 이용한 실험을 통해 살펴보았다. 선형성을 판단하는 보조 수단으로서 도움이 된다는 것을 알 수 있었으나 해결해야 할 문제도 있다. 한계점과 해결해야 할 문제를 지적하기에 앞서, 일반화선형모형에서 $g(\mu) = F(\boldsymbol{x})$를 만족하는 함수 F를 비모수적인 방법으로 추정하기 위해 부스팅 기법을 이용하였는데, 이 때 분류문제가 아니라 회귀문제임을 강조할 필요가 있다. 이항형 자료의 경우에도 반응변수 y의 적합값 $\{0, 1\}$을 예측하는 것이 아니라 $F(\boldsymbol{x})$를 추정하는 문제이므로 회귀문제이다. 분류문제에서 부스팅 기법의 성능에 대한 연구는 많이 이루어졌으나 상대적으로 회귀문제에서는 그 성능과 방법에 대한 연구가 많지 않은데, 예를 들어 부스팅 기법에서 관측값에 대한 가중값을 달리 하면서 반복수행하는 회수 (또는 결합하는 나무모형의 개수) M의 최적값을 정하는 문제도 간단하지 않다. 분류문제가 아닌 회귀문제에서 비록 깊이가 1인 나무모형 (stump)을 기본학습기 (weak learner)로 쓰더라도 과적합 (overfitting)이 발생하는데, 이 때 M의 최적값을 정해주는 문제가 중요하다. (앞의 모의실험에서 검증자료 (test data)를 이용하여 M의 최적값을 정하였으나 실제 자료에 적용할 때 검증자료가 따로 존재하지 않는 경우에는 교차타당성이나 붓스트랩에 의해 M의 최적값을 정할 수도 있다.) 그 외에 나무모형의 적정 깊이를 정하는 문제 (상호작용이 있을 때 깊이 1의 나무

모형은 적절하지 않음) 등이 있다. 하지만 $\hat{\mu}_{BOOST}$를 구하는 과정에 대한 자세한 설명은 생략하였는데, 그 이유는 이 연구의 진단그림에서 비모수적 회귀에 의한 추정값 $\hat{\mu}_{NP}$으로 부스팅 기법에 의한 추정값 $\hat{\mu}_{BOOST}$를 사용하였으나 다른 비모수적 회귀 방법도 얼마든지 가능하기 때문이다.

이 연구의 진단그림은 연속형 설명변수가 있고 표본 크기가 어느 정도 클 때 유용하다. 설명변수가 모두 범주형인 경우 점들이 겹쳐서 표시되기 때문에 진단그림으로 선형성 여부를 판단하기 어려운 한계가 있다. 하지만 이 경우 그룹화를 이용하여 모형의 적합성을 판단할 수 있는 그림을 그릴 수 있으므로 새로운 진단그림을 필요로 하지 않는다.

반응변수의 분포가 포아송 분포인 경우와, 일반화선형모형뿐만 아니라 Cox 비례위험모형의 적합도 진단에 적용하는 것도 앞으로 연구해볼 가치가 있다.

참고문헌

Cheng, K. F. and Wu, J. W. (1994). Testing goodness of fit for a parametric family of link functions. *Journal of the American Statistical Association,* **89**, 657–664.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall/CRC, London.

Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences,* **55,** 119–139.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics,* **28,** 337–374.

Kim, J. H. (2003). Assessing practical significance of the proportional odds assumption. *Statistics & Probability Letters,* **65,** 233–239.

Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association,* **79,** 61–71.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics,* **31,** 172–181.

Su, J. Q. and Wei, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association,* **86,** 420–426.

Therneau, T. M. and Atkinson, B. (2006). rpart: Recursive Partitioning. R package version 3.1-33. S-PLUS 6.x original at http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm.

[2007년 7월 접수, 2007년 10월 채택]

A Graphical Method of Checking the Adequacy of Linear Systematic Component in Generalized Linear Models†

Ji-Hyun Kim[1)]

Abstract

A graphical method of checking the adequacy of a generalized linear model is proposed. The graph helps to assess the assumption that the link function of mean can be expressed as a linear combination of explanatory variables in the generalized linear model. For the graph the boosting technique is applied to estimate nonparametrically the relationship between the link function of the mean and the explanatory variables, though any other nonparametric regression methods can be applied. Through simulation studies with normal and binary data, the effectiveness of the graph is demonstrated. And we list some limitations and technical details of the graph.

Keywords: Boosting; nonparametric regression; bootstrap confidence intervals.

† This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-015-C00086).

1) Professor, Department of Statistics and Actuarial Science, Soongsil University, Dongjak-Ku Sangdo-Dong, Seoul 156-743, Korea. E-mail: jxk61@ssu.ac.kr