

범주형 자료분석 개론 3.4~3.5

2021210088 허지혜

3.4.1 가능도함수를 이용한 왈드, 가능도비, 스코어 추정

표본이 큰 경우 GLM의 ML 추정량들은 근사적으로 정규분포를 따른다.

앞서 제시된 추론 방법들을 설명 변수가 하나인 GLM의 관점에서 다시 검토해보자.

1. 가능도함수를 이용한 왈드 추정

설명변수가 한 개인 GLM에서 설명변수 x 가 반응변수에 영향을 미치지 않음을 나타내는 귀무가설을 검정하기 위한 왈드 검정통계량은 $z = \hat{\beta}/SE$ 이고 SE 는 제약이 없을 때의 $\hat{\beta}$ 표준오차이다. 귀무가설하에서 z 는 근사적으로 표준정규분포를 따른다. z^2 은 근사적으로 자유도가 1인 카이제곱분포를 따른다.

2. 가능도함수를 이용한 가능도비 추정

l_1 : 완전모형 하에서의 가능도함수의 최댓값

l_0 : 귀무가설 하에서의 가능도함수의 최댓값 일때 가능도비 검정통계량은 다음과 같다.

$$2\log(l_1/l_0) = [2\log(l_1) - \log(l_0)] = 2(L_1 - L_0)$$

L_1 : 완전모형 하에서의 로그 가능도함수의 최댓값

L_0 : 귀무가설 하에서의 로그 가능도함수의 최댓값

귀무가설하에서 통계량은 근사적으로 자유도가 1인 카이제곱분포를 따른다.

3.4.1 가능도함수를 이용한 왈드, 가능도비, 스코어 추정

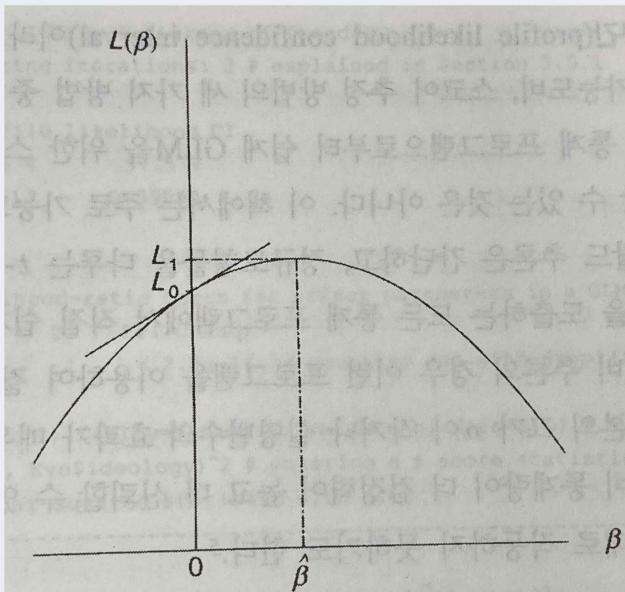


그림 3.5

귀무가설에 대한 GLM 검정에서 사용된 로그 가능도함수 $L(\beta)$.
왈드 통계량은 $\hat{\beta}$ 과 $\hat{\beta}$ 값에서의 가능도함수 $L(\beta)$ 의 곡률을 사용한다.
가능도비 검정은 $(L_1 - L_0)$ 의 2배 값을 이용한다.
스코어 검정은 $\beta = 0$ 에서 $L(\beta)$ 에 그은 접선의 기울기를 사용한다.

이항 로지스틱 회귀모형이나 포아송 로그 선형모형을 포함해 몇몇 GLM에 대한 로그 가능함수는 오목한 모양을 가진다.

ML 추정량 $\hat{\beta}$ 은 로그 가능도함수가 최댓값을 갖는 점이다.

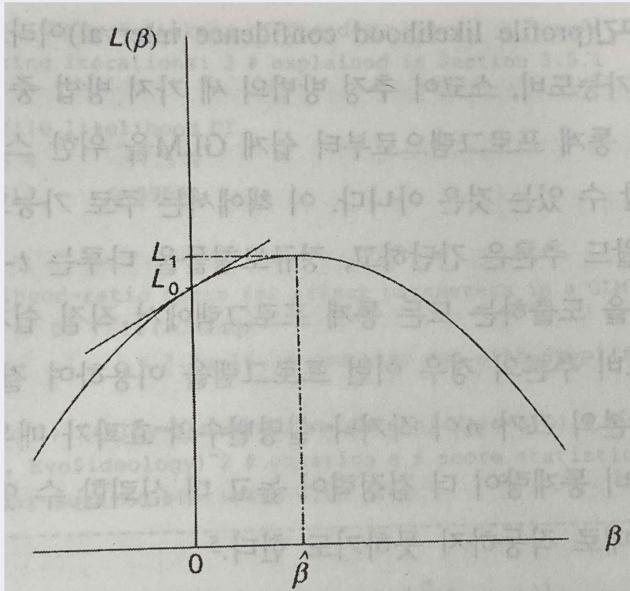
가능도비 통계량인 $2(L_1 - L_0)$ 은 $\hat{\beta}$ 에서의 가능도함수값과 $\beta = 0$ 에서의 가능도함수값 간 수직 거리의 두 배가 된다.

왈드 검정법은 ML 추정값 $\hat{\beta}$ 에서의 가능도함수값 $L(\beta)$ 만을 사용한다.

표본크기가 커질수록 $L(\beta)$ 의 곡률이 증가하고 SE 값이 작아진다.

곡률이 크다는 것은 $L(\beta)$ 가 β 에서 멀어질수록 로그 가능도함수값이 더 빠르게 떨어지며 가능한 β 값들의 범위가 더 좁아짐을 의미한다.

3.4.1 가능도함수를 이용한 왈드, 가능도비, 스코어 추정



스코어 검정은 귀무가설의 β 값인 0에서의 로그 가능도함수 모양에만 영향을 받는다.

이 검정은 귀무가설에서 가능도함수 $L(\beta)$ 에 접하는 선을 그릴 때 그 선의 기울기를 사용한다.

이 기울기 값은 $\hat{\beta}$ 가 귀무가설의 값으로부터 멀어질수록 그 절댓값이 커지게 된다.

스코어 통계량은 이 기울기값을 귀무가설의 β 값을 이용해 계산한 SE로 나눈 비로 정의된다.

스코어 통계량의 제곱은 근사적으로 자유도가 1을 갖는 카이제곱분포를 따른다.

이 책에서는 스코어 통계량의 일반적인 형태는 다루지 않는다.

하지만 범주형 자료에서 주로 사용되는 검정통계량은 이러한 스코어 통계량의 형태로 주어진다.

독립성검정을 위한 Pearson χ^2 통계량과 순서형 상관계수 검정통계량 M^2 도 스코어 통계량의 한 형태이다.

3.4.1 가능도함수를 이용한 왈드, 가능도비, 스코어 추정

왈드, 가능도비, 스코어 추정 방법 각각에 대응하는 신뢰구간을 구할 수 있다.

β 에 대한 95% 신뢰구간은 $H_0: \beta = \beta_0$ 에 대한 검정에서 P-값이 0.05가 넘는 모든 β_0 값들을 포함한다.

예로, H_0 에 대한 왈드 통계량은 $z = (\check{\beta} - \beta_0)/SE$ 이며 신뢰구간은 $\hat{\beta} \pm Z_{\alpha/2}(SE)$ 이다.

기능도비에 기초해서 구한 β 에 대한 신뢰구간은 **프로파일 가능도 신뢰구간**이라고 부른다.

왈드, 가능도비, 스코어 추정 방법의 세 가지 방법 중 어떤 방법을 주로 써야 할까 ?

이 책에서는 주로 가능도비와 왈드 추론 방법을 다룬다.

왈드 추론은 간단하고, 정규모형들을 다루는 t-방법과 비슷하며, ML 추정값과 SE 값을 도출하는 통계 프로그램에서 직접 쉽게 결과를 얻을 수 있다.

그러나 가능도비 추론으로 결과를 직접 얻기가 쉽지 않다.

그러나 표본의 크기가 작거나 설명변수의 효과가 매우 큰 경우에는

왈드 통계량보다 가능도비 통계량이 더 검정력이 높고 신뢰할 수 있다.

3.4.2 예제 : 정치성향과 진화에 대한 믿음

2016년 일반사회조사에서는 “인류는 우리가 알고 있는 것처럼 초기의 동물 종에서 진화하였다. 맞는가? 틀린가?”라는 질문이 있다.

이와 같은 질문에 대한 답과 정치성향은 관련성이 있을까?

y : 진화에 대한 의견(1 = 맞다, 0 = 틀리다)

x : 정치성향(1 = 매우 보수적, 2 = 보수적, 3 = 약간 보수적, 4 = 중간, 5 = 약간 진보적, 6 = 진보적, 7 = 매우진보)

```
> Evo <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Evolution.dat", header=TRUE)
> Evo
  ideology true false
  1       1     11    37
  2       2     46   104
  3       3     70    72
  4       4    241   214
  5       5     78    36
  6       6     89    24
  7       7     36     6
> n <- Evo$true + Evo$false
> fit <- glm(true/n ~ ideology, family=binomial, weights= n, data=Evo)
> summary(fit)

Call:
glm(formula = true/n ~ ideology, family = binomial, data = Evo,
     weights = n)

Deviance Residuals:
      1      2      3      4      5      6      7 
 0.1430 -0.2697  1.4614 -1.0791   0.2922   0.4471   0.2035 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.75658   0.20500 -8.569   <2e-16 ***
ideology     0.49422   0.05092  9.706   <2e-16 ***
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 113.20  on 6  degrees of freedom
Residual deviance:  3.72  on 5  degrees of freedom
AIC: 42.332

Number of Fisher Scoring iterations: 3

> confint(fit)
Waiting for profiling to be done...
              2.5 %    97.5 %
(Intercept) -2.165294 -1.3609733
ideology     0.396166  0.5959414
> library(car)
> Anova(fit)
Analysis of Deviance Table (Type II tests)

Response: true/n
          LR Chisq Df Pr(>Chisq)    
ideology 109.48  1  < 2.2e-16 ***
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> library(statmod)
> fit0 <- glm(true/n ~1, family=binomial, weights=n, data=Evo)
> glm.scoretest(fit0, Evo$ideology)^2
[1] 104.101
```

3.4.2 예제 : 정치성향과 진화에 대한 믿음

```
> Evo <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Evolution.dat", header=TRUE)
> Evo
  ideology true false
1          1    11    37
2          2    46   104
3          3    70    72
4          4   241   214
5          5    78    36
6          6    89    24
7          7    36     6
> n <- Evo$true + Evo$false
> fit <- glm(true/n ~ ideology, family=binomial, weights= n, data=Evo)
> summary(fit)

Call:
glm(formula = true/n ~ ideology, family = binomial, data = Evo,
     weights = n)

Deviance Residuals:
      1      2      3      4      5      6      7 
 0.1430 -0.2697  1.4614 -1.0791  0.2922  0.4471  0.2035 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.75658   0.20500 -8.569   <2e-16 ***
ideology     0.49422   0.05092  9.706   <2e-16 ***
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 113.20 on 6 degrees of freedom
Residual deviance:  3.72 on 5 degrees of freedom
AIC: 42.332

Number of Fisher Scoring iterations: 3
```

로지스틱 회귀모형의 ML 적합 결과는 $\text{logit}[\hat{P}(y=1)] = -1.757 + 0.494x$ 이고
정치성향의 효과를 나타내는 모수는 $\hat{\beta} = 0.494$ 이고 $SE = 0.051$ 이다.

$$z = \frac{\hat{\beta}}{SE} = \frac{0.494}{0.051} = 9.71$$

귀무가설에 대한 왈드 검정통계량은 다음과 같이 주어지고 정규분포를 따르며
 $Z^2 = 96.2$ 은 자유도가 1인 카이제곱 분포를 따른다.

3.4.2 예제 : 정치성향과 진화에 대한 믿음

이 결과는 진화를 믿을 확률이 정치성향이 진보적일수록 높아진다는 아주 강한 증거를 보여 주고 있다.
이 모형을 $\beta = 0$ 을 가정한 간단한 모형과 비교한 가능성도비 검정에서도 유사하게 강한 증거를 얻을 수 있다.

R에서 car 패키지에 있는 Anova 함수를 적용해 검정을 실시할 수 있다.
이 경우, 카이제곱통계량은 109.48의 값을 가지며 df = 1 이다.

```
> confint(fit)
Waiting for profiling to be done...
      2.5 %    97.5 %
(Intercept) -2.165294 -1.3609733
ideology     0.396166  0.5959414
> library(car)
> Anova(fit)
Analysis of Deviance Table (Type II tests)

Response: true/n
          LR Chisq Df Pr(>Chisq)
ideology 109.48  1 < 2.2e-16 ***
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> library(statmod)
> fit0 <- glm(true/n ~1, family=binomial, weights=n, data=Evo)
> glm.scoretest(fit0, Evo$ideology)^2
[1] 104.101
```

β 에 대한 95% 왈드 신뢰구간은 $\hat{\beta} \pm 1.96(SE)$ 이고 $0.494 \pm 1.96(0.051) = (0.394, 0.594)$
R의 출력 결과값에서 β 에 대한 프로파일 가능성도 95% 신뢰구간은 $(0.396, 0.596)$ 이다.

z 스코어 통계량은 R의 statmode 패키지에 있는 glm.scoretest 함수를 사용해
구할 수 있는데 귀무가설 하의 모형에 정치성향 변수를 추가하면 된다.
z 스코어 통계량의 제곱은 df = 1인 카이제곱분포를 따르며 104.101이다.

3.4.3 GLM의 이탈도 통계량

M : 관심 있는 모형

L_M : 모형 M에서 얻은 로그 가능성함수의 최댓값

L_S : 가능한 모형 중에서 가장 복잡한 형태의 모형하에서의 로그 가능성함수의 최댓값

=> 이 모형은 각 관측값에 대해 모수를 갖게 되므로 완벽하게 자료를 적합시킨다.

=> 이 모형을 **포화모형**이라고 한다.

예시) 7X2 분할표에서 로지스틱 회귀모형 M을 적합시켰다.

	ideology	true	false
1	1	11	37
2	2	46	104
3	3	70	72
4	4	241	214
5	5	78	36
6	6	89	24
7	7	36	6

이 모형은 정치성향의 7가지 수준에 따라 진화를 믿을 확률이 어떻게 변하는지 나타내기 위해 두 개의 모수를 사용한다.

반면, 이 자료에 대한 포화모형의 경우에는 정치성향의 7개 수준별로 각 이항 관측값에 대하여 서로 다른 모수를 가정한다.

매우 보수적인 사람들의 경우 : $P(Y=1) = \pi_1$

보수적인 사람들의 경우 : π_2, \dots

매우 진보적인 사람들의 경우 : π_7

포화모형에서 π_i 에 대한 ML 추정값은 i번째 정치성향 수준에서 진화를 믿을 표본확률값이다.

3.4.3 GLM의 이탈도 통계량

포화모형은 추가적으로 모수들을 더 많이 포함하기 때문에 이 모형하에서 구한
로그 가능도함수의 최댓값 L_S 는 적어도 더 단순한 모형 M에서 구한 로그 가능도함수 L_M 보다 큰 값을 갖는다.

GLM의 이탈도 통계량 = $2[L_S - L_M]$

이탈도 통계량은 포화모형 S와 관심 있는 모형 M을 비교하기 위한 가능도비 통계량으로
포화모형에 있는 모수들 중에서 모형 M에 포함되지 않은 모수들이 모두 0이라는 귀무가설을 검정하기 위한
통계량이다.

GLM 모형에서 이탈도 통계량은 근사적으로 카이제곱분포를 따른다.

예시) 5.2.1절에서 설명변수의 수준 수가 고정되어 있고 성공도수와 실패도수가 상대적으로 큰 값을 갖는 이항
GLM 모형에 대해서 이러한 사실을 살펴볼 것이다.

이런 경우, 이탈도 통계량은 모형의 적합성을 검정하는데 이것은 모형에 포함되지 않은 모수들이 모두 0이라는
귀무가설을 검정하기 때문이다.

잔차 자유도는 전체 관측값의 수에서 모형의 모수 수를 뺀 값이다.

P-값은 카이제곱분포에서 관측된 검정통계량의 오른쪽 꼬리부분의 확률이다.

검정통계량 값이 크고 P-값이 작을수록 모형의 적합결여에 대한 강한 증거가 된다.

3.4.3 GLM의 이탈도 통계량

R에서 `glm` 함수로 일반화선형모형을 적합시킬 경우, 잔차 이탈도는 적합된 모형의 이탈도를 나타내며, 영이탈도는 상수항만 가지는 귀무가설 하에서의 모형의 이탈도를 나타낸다.
잔차의 정치성향에 대한 자료에서 로지스틱 회귀모형은 두 개의 모수를 사용해 7개의 이항 관측값을 설명하고 있다.

앞선 예제에서 계산된 잔차 이탈도는 3.72이고 $df = 7 - 2 = 5$ 이다.
이 모형이 가정하는 귀무가설에 대한 검정 결과 P-값은 0.59이다.
즉, 모형이 자료를 잘 적합한다고 할 수 있다.

3.4.4 이탈도 통계량을 이용한 모형 비교

두 모형에 대해서 한 모형(M_0)이 다른 모형(M_1)의 특별한 경우라고 가정해 보자.

즉, M_0 가 M_1 의 내포모형이라고 하자. 이때 두 모형을 비교하기 위해 이탈도 통계량을 사용할 수 있다.

정규분포를 따르는 반응변수들에 대한 모형들을 비교할 때 F검정을 사용하듯 GLM에서는 가능도비 검정을 사용한다.

모형의 비교를 위한 F검정은 자료의 분산을 나타내는 제곱합을 분해한다.

분산을 분해시키기 위한 이 **분산분석**을 GLM에 대한 **이탈도 통계량분석**으로 일반화시킬 수 있다.

좀 더 복잡한 모형이 만족될 때, 더 간단한 모형이 만족되는지를 검정하기 위한 가능도비 검정통계량은 다음과 같으며 모형들의 이탈도 통계량을 통해서 두 모형을 비교 가능하다.

$$2(L_1 - L_0) = 2(L_s - L_0) - 2(L_s - L_1) = \text{이탈도 통계량}_s^2 - \text{이탈도 통계량}_1^2$$

모형 M_0 가 M_1 에 비해 적합이 잘 되지 않는 경우에 이 검정통계량의 값은 큰 값을 가진다.

대표본의 경우 이 통계량은 근사적으로 카이제곱분포를 따르게 되며 자유도는 각 모형들의 자유도의 차이와 같다.

이 자유도의 값은 모형 M_1 에는 포함되어 있지만 모형 M_0 에는 포함되지 않은 모수들의 수와 같다.

검정통계량 값이 크고 P-값이 작을수록 모형 M_0 가 M_1 보다 잘 적합되지 않는다는 것을 의미한다.

3.4.2절에 진화와 정치성향 자료의 경우) 계산된 잔차 이탈도 = 3.72, df = 7 - 2 = 5이다.

정치성향의 효과를 고려하지 않은 더 간단한 모형에서는 자유도는 6이고 이탈도통계량은 113.20이다.

두 이탈도 통계량의 차이는 df = 1 에서 104.48인데 이것은 정확히 $\beta = 0$ 을 검정하기 위한 가능도비 검정통계량이다. 일반적으로 모형에 대한 영이탈도와 잔차 이탈도의 차이는 모형에 있는 β 효과항이 0이라는 가설을 검정하기 위한 가능도비 통계량과 같다.

3.4.5 관측값과 모형적합값을 비교하는 잔차

관측값 i 에 대해서 관측도수와 적합도수의 차이를 나타내는 잔차 $y_i - \hat{\mu}_i$ 는 쓰임이 제한적이다.
예로, 포아송 표본추출에 대해서 도수의 표준편차가 $\sqrt{\mu_i}$ 이므로 μ_i 가 커질수록 그 차이도 커지는 경향을 보이게 된다.
Pearson 잔차는 이 차이를 표준화한 것으로 다음과 같이 정의한다.

$$\text{Pearson 잔차} = e_i = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\mu}_i}}$$

포아송 GLM에 대해서 $\text{var}(y_i) = \mu_i$ 이므로 Pearson 잔차는 다음과 같다.

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

이것은 잔차를 포아송 표준편차의 추정값으로 나누어 표준화시킨 것이다.

e_i 를 Pearson 잔차라고 부르는 이유는 $\sum e_i^2 = \sum (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i$ 이기 때문이다.

GLM이 이차원 분할표에서 칸들 간의 독립성에 대응되는 모형일 때 이것은 독립성을 검정하는 Pearson 카이제곱 통계량 X^2 이 된다. 그러므로 X^2 는 각각그이 관측값에 적합 결여를 나타내는 항으로 분해된다.

이탈도 잔차라고 불리는 이탈도 통계량의 성분은 적합결여를 나타내는 또 다른 측도이다.

3.4.5 관측값과 모형적합값을 비교하는 잔차

Pearson 잔차의 분모는 y_i 의 변동을 설명하지만 $\hat{\mu}_i$ 의 변동을 설명하지는 못한다.
표준화잔차는 $y_i - \hat{\mu}_i$ 를 표준오차의 추정값으로 나눈 값이다.

$$\text{표준화잔차} = (y_i - \hat{\mu}_i) / SE$$

표준화잔차는 y_i 와 $\hat{\mu}_i$ 의 변동 모두를 설명하기 때문에 Pearson 잔차나 이탈도 잔차보다 선호되는 경향이 있다.
표준화잔차를 사용하면 $y_i - \hat{\mu}_i$ 가 언제 큰 값을 갖는지 쉽게 판단할 수 있다.
 $\hat{\mu}_i$ 가 크면 표준화잔차는 근사적으로 표준정규분포를 따른다.

비록 표본이 큰 경우는 몇몇 표준화잔차값 중에서 우연히 2나 3보다 큰 수가 나올 수도 있지만 이런 경우에는 주의 깊게 살펴볼 필요가 있다.

뒷 절에서는 로지스틱 회귀분석에서 표준화 잔차를 사용하고, 표준화 잔차가 Pearson 잔차나 이탈도 잔차보다 잔차 자유도를 더 잘 반영하고 있다는 것을 보일 것이다.

3.4.5 관측값과 모형적합값을 비교하는 잔차

다음의 결과값은 Evolution 자료 파일에서 정치성향, 진화를 믿는 사람의 도수(true), 진화를 믿지 않는 사람의 도수(false), 이항포본의 크기(n), 정치성향별 진화를 믿을 확률에 대한 표본비율, 로지스틱 모형의 비율 적합값, 표준화잔차값을 보여준다.
표본비율값은 로지스틱 모형의 비율 적합값에 가깝다.
어떤 표준화잔차도 적합결여를 나타내지 않고 있는데, 이는 잔차 이탈도값으로부터 모형이 잘 적합되고 있다는 것.

```
> cbind(Evo$ideology, Evo$true, Evo$false, n,Evo$true/n,fitted(fit),rstandar$  
      n  
1 1 11 37 48 0.2291667 0.2205679 0.1611162  
2 2 46 104 150 0.3066667 0.3168813 -0.3515386  
3 3 70 72 142 0.4929577 0.4319445 1.6480176  
4 4 241 214 455 0.5296703 0.5548525 -1.4995488  
5 5 78 36 114 0.6842105 0.6713982 0.3248519  
6 6 89 24 113 0.7876106 0.7700750 0.5413625  
7 7 36 6 42 0.8571429 0.8459201 0.2206605  
> restandard(fit, type="person")  
Error in restandard(fit, type = "person") :  
  할수 "restandard"을 찾을 수 없습니다  
> rstandard(fit, type="person")  
Error in match.arg(type) :  
  'arg' should be one of "deviance", "pearson"  
> rstandard(fit, type="pearson")  
    1         2         3         4         5         6         7  
 0.1611162 -0.3515386 1.6480176 -1.4995488 0.3248519 0.5413625 0.2206605  
> |
```

cbind(Evo\$ideology, Evo\$true, Evo\$false, n,Evo\$true/n,fitted(fit),rstandard(fit,type="person"))
rstandard(fit, type="pearson")

회귀모형에서 사용되는 다른 진단 방법들도 GLM의 적합도를 검토하는 데 도움이 된다.
예를 들어, 모형의 적합에서 하나의 관측값의 영향력을 측정하기 위해 그 관측값을 제거한 후 모형을 다시 적합하여 기존 모형의 적합 결과와 비교할 수 있다.

3.5 일반화선형모형의 적합

통계학에서는 가능도함수를 최대화시키는 ML 추정값을 구하는 과정에서 기본적인 미분적분학을 사용한다. 로그 가능도함수를 다양한 모수들에 대하여 미분한 후, 미분한 값이 0로 놓은 “가능도 방정식”을 유도 후 이 가능도 방정식을 풀어서 ML 추정값을 구할 수 있다.

특별한 경우를 제외하면, GLM에 대한 가능도 방정식은 닫힌 해를 가 가지고 있지 않다. 통계 프로그램에서는 알고리즘을 사용하여 가능도 방정식을 풀고 모형을 적합한다.

3.5.1 Fisher 스코어 알고리즘

먼저 가능도함수를 최대화하는 모수값에 대한 초기값을 추측한다.

위 알고리즘이 성공적으로 수행되면 초기값은 점점 ML 추정값에 가까워지게 된다.

Fisher 스코어 알고리즘은 Fisher가 이항 회귀모형을 적합하기 위해 처음으로 제안한 방법이다.

알고리즘 내 각 반복 과정은 일종의 가중최소제곱법을 반영하였다.

동질하지 않은 γ 의 분산을 고려하여 보통의 최소제곱법을 일반화시킨 것이다.

분산이 작은 관측값에 더 많은 가중치를 준다.

각 반복마다 ML 추정값의 근사값이 변하고 이에 따른 분산 추정값도 변하기 때문에 가중치를 계속 변한다.

이러한 GLM에 대한 ML 추정을 반복재가중최소제곱 추정이라고 부른다.

위를 반복적이라고 하는 이유는 알고리즘에서 로그 가능도값이 더 이상 증가하지 않을 때까지 위 과정을 반복하기 때문이다.

근사가 성공적이면 단지 몇 번의 반복 과정만으로 ML 추정값에 빠르게 수렴한다.

예시) 3.4.2절에 정치성향과 진화에 대한 믿음의 연관성을 나타낸 로지스틱 회귀모형

=> Fisher 스코어 알고리즘을 수행해 3번의 반복 과정을 거쳐 수렴

3.5.1 Fisher 스코어 알고리즘

이항 로자스틱 회귀모형과 포아송 로그선형 모형에 대해서

Fisher 스코어 알고리즘이 **Newton-Raphson 알고리즘**과 일치하게 된다.

Newton-Raphson 알고리즘은 추정된 초기값 주위에서 로그 가능도함수를 오목한 포물선 모양을 갖는 간단한 다항식함수로 근사시킨다.

이 다항식은 초기값에서 로그 가능도함수와 동일한 기울기와 굴곡을 가진다.

이 다항식으로부터 최댓값의 위치를 쉽게 구할 수 있다.

그 위치에서부터 ML 추정값에 대한 두 번째 추정값을 찾을 수 있다.

다시 다른 오목한 포물선 함수를 사용하여 두 번째 추정값의 주위에서 로그 가능도함수를 근사한 후에 이 함수를 사용해 세 번째 추정값을 찾는다

위 과정을 반복해 ML 추정량을 구하게 된다.

3.5.2 일반화선형모형에 대한 베이지안 방법

일반화선형모형에서 효과를 나타내는 모수는 실수 전체값을 가질 수 있다.

$\{\beta_j\}$ 에 대한 베이지안 추론에서는 모수가 다변량 정규분포를 따른다고 가정함으로써 사전분포에 대한 유연성을 줄 수 있다.

정보가 없는 사전분포의 간단한 형태는 모형 내 각각의 모수가 서로 독립이고 정규분포를 따른다고 가정한다. 모형의 각 모수에 대한 정보가 없는 사전분포의 표준편차가 동일할 경우, 양적인 설명변수를 표준화함으로써 사전효과의 크기를 동일한 단위로 나타낼 수 있다. 설명변수를 표준화시키지 않을 경우에는 측정단위를 고려해야 한다.

예시) x_j = 시간을 연 단위에서 월 단위로 바꿀 경우, 새로운 β_j 는 x 가 연도 단위로 측정되었을 때의 β_j 의 값의 $1/12$ 이 되므로, 정규 사전의 새로운 표준편차 값에도 $1/12$ 를 곱해주어야 한다.

몬테칼로 시뮬레이션 방법을 사용하는 통계 프로그램을 이용하면 GLM에 대한 베이지안 방법을 계산할 수 있다. 위 예는 뒤에서 살펴볼 예정이다.

3.5.3 GLM의 장점

GLM이론은 1970년대 중반 연속형 반응변수와 범주형 반응변수들에 대한 중요한 모형들을 통합하는 역할을 하면서 발전해 왔다.

밑의 표는 통계분석에서 실제로 많이 쓰이는 GLM을 열거하였다.

GLM의 장점은 모형적합 알고리즘이 Fisher 스코어 알고리즘이 어떠한 GLM에 대해서도 동일하다는 것이다.

즉, 랜덤성분의 분포나 연결함수와는 상관없이 동일하게 알고리즘을 적용할 수 있다.

=> GLM 소프트웨어들을 이용해 다양한 종류의 모형들을 적합시킬 수 있다.

랜덤 요소	연결함수	체계적 성분	모형	장(chapter)
정규분포	항등	연속형	회귀모형	
정규분포	항등	범주형	분산분석	
정규분포	항등	혼합형	공분산분석	
이항분포	로짓	혼합형	로지스틱 회귀	4-5, 8-10
다항분포	로짓	혼합형	다항 반응	6, 8-10
포아송 분포	로그	혼합형	로그 선형	7