

# 범주형 자료분석 개론

2021210088 허지혜

## 4.5 로지스틱 회귀모형의 효과에 대한 요약

### 4.5.1 확률에 기초한 해석

예측변수  $x_j$  의 효과를 설명할 수 있는 방법은 다른 예측변수들의 값을 표본 평균값으로 고정시킨 후  $x_j$  의 가장 작은 값과 가장 큰 값에서  $\hat{P}(Y = 1)$  를 구하는 것이다.

이 두  $\hat{P}(Y = 1)$  값이나 두 값의 차이를 통해 예측변수  $x_j$  의 효과를 알 수 있다.

그러나 연속형 설명변수인 경우, 이렇게 구한 값들이 타당한지 여부가 다른 모든 예측변수들의 값이 표본평균 근처에 있을 때에  $x_j$  가 극단값을 갖는지에 따라 좌우된다는 것이다.  $x_j$  의 이상점이 존재하는 경우에도 이 값들은 이상점에 의해 영향을 받으므로  $x_j$  의 사분위수를 사용하는 것이 적절하다.  
사분위수를 사용할 경우 예측변수의  $x_j$  의 효과는  $x_j$  값들의 전체 범위중 중간 범위의  $\hat{P}(Y = 1)$  값들의 변화를 설명하게 된다.

변수	추정값	표준오차	비교	확률의 변화량
너비(x)	0.478	0.104	(max,min) at $\bar{c}_4$ (UQ, LQ) at $c_4$	$0.84 = 0.985 - 0.142$ $0.29 = 0.803 - 0.516$
색깔( $c_4$ )	-1.300	0.526	(0,1) at $\bar{x}$	$0.31 = 0.710 - 0.401$

## 4.5 로지스틱 회귀모형의 효과에 대한 요약

### 4.5.1 확률에 기초한 해석

변수	추정값	표준오차	비교	확률의 변화량
너비(x)	0.478	0.104	(max,min) at $\underline{c_4}$ (UQ, LQ) at $\bar{c_4}$	$0.84 = 0.985 - 0.142$ $0.29 = 0.803 - 0.516$
색깔( $c_4$ )	-1.300	0.526	(0,1) at $\bar{x}$	$0.31 = 0.710 - 0.401$

너비와 색깔을 나타내는 예측변수를 가지고 있는 암참게 자료에서 예측식은 다음과 같다.

$$\begin{aligned} \text{너비의 표본평균값에서 } c_4 = 1 &\Rightarrow \hat{P}(Y = 1) = 0.40 \\ c_4 = 0 &\Rightarrow \hat{P}(Y = 1) = 0.71 \end{aligned}$$

어두운 색의 참게와 다른 색의 참게를 구분하는 색깔의 효과가 실제로 큼을 알 수 있다.

$c_4$ 이 표본평균값을 가질 때,  $\hat{P}(Y = 1)$  는 너비가 최솟값에서 최댓값으로 변할 때 0.14에서 0.98까지 변하고, 너비값의 중앙 50%에 해당하는 범위에서 0.52에서 0.80까지 변하므로 너비의 효과도 강하다는 것을 알 수 있다.

$c_4$ 는 0과 1의 값만 가지므로  $c_4$ 의 평균값에서 효과를 살펴보는 대신에  $c_4$ 의 각 값에서 따로 효과를 살펴보는 것이 나을 것이다.

R에서 이와 같은 효과의 요약값을 얻기 위한 방법은 다음과 같다.

## 4.5 로지스틱 회귀모형의 효과에 대한 요약

```
> Crabs$c4 <- ifelse(Crabs$color == 4,1,0)
> fit3 <- glm(y ~ width + c4, family=binomial, data=Crabs)
> predict(fit3, data.frame(c4=1, width=mean(Crabs$width)), type="response")
  1
0.4006293
> predict(fit3, data.frame(c4=0, width=mean(Crabs$width)), type="response")
  1
0.7104701
> predict(fit3, data.frame(c4=mean(Crabs$c4), width=quantile(Crabs$width)), 
+   0%      25%      50%      75%      100%
0.1416394 0.5158264 0.6541188 0.8025564 0.9848731
: 1
```

## 4.5.2 주변 효과와 그 평균

앞에서 상대적으로 작은 양적 예측변수의 변화가 확률에 미치는 변화를 근사시키기 위해 직선의 기울기를 사용하였다. 이와 같은 간단한 해석 방법은 여러 예측변수가 있는 경우에도 적용된다.

$\hat{P}(Y=1) = \hat{\pi}$ 인 설명변수들의 조합을 고려해 보자. 다른 설명변수들을 고정했을 때,  $x_j$  가 한 단위 증가할 때마다 근사적으로  $\hat{\pi}$  는  $\beta_j \hat{\pi}(1 - \hat{\pi})$  만큼 변화한다.

$$\text{logit}[\hat{P}(Y=1)] = -11.68 + 0.487x - 1.300c_4$$

예) 참게자료에서 예측변수  $x$  : 너비,  $c_4$  : 어두운 색깔의 참게일 경우 1, 아니면 0으로 정의되는 지시변수인 경우 너비  $x$ 의 효과에 대한 추정값은  $\hat{\beta}_4 = 0.478$ 이다.

$\hat{\pi} = 0.50$ 일 때,  $x$ 가 1cm 증가하면  $\hat{\pi}$  의 근사적인 효과는  $0.478(0.50)(0.50) = 0.12$  이다.  
이 효과는 무시할 수 없을 정도로 큰 편인데 너비 1cm의 증가량은 표준오차 2.1cm의 절반도 안되는 값이기 때문이다.

설명변수의 효과를 설명하는 변화율은 값에 따라 달라진다. 이 효과의 전체 요약값은  $n$ 개의 설명변수의 표본값들로부터 구한 변화율의 평균값이다. 일부 통계프로그램에서는 이 측정값을 평균주변효과라고 부른다. 이항설명변수의 경우, 두 범주에 대한  $P(Y=1)$ 의 추정값의 차이의 평균을 구할 수 있는데 이 값을 이산변화량이라고 부른다. 너비의 색깔을 설명변수로 가지며 색깔이 이항변수인 로지스틱 모형의 예를 살펴보자.

## 4.5.2 주변 효과와 그 평균

```
> fit3 <- glm(y ~ width + c4, family=binomial, data=Crabs)
> library(mfx)
필요한 패키지를 로딩중입니다: sandwich
필요한 패키지를 로딩중입니다: lmtest
필요한 패키지를 로딩중입니다: zoo

다음의 패키지를 부착합니다: 'zoo'

The following objects are masked from 'package:base':

  as.Date, as.Date.numeric

필요한 패키지를 로딩중입니다: MASS
필요한 패키지를 로딩중입니다: betareg
경고메시지 (2):
1: 패키지 'mfx'는 R 버전 4.0.5에서 작성되었습니다
2: 패키지 'sandwich'는 R 버전 4.0.5에서 작성되었습니다
3: 패키지 'betareg'는 R 버전 4.0.5에서 작성되었습니다
> logitmfx(fit3,atmean=FALSE, data=Crabs)
Call:
logitmfx(formula = fit3, data = Crabs, atmean = FALSE)

Marginal Effects:
      dF/dx Std. Err.      z   P>|z|
width  0.087483  0.024472  3.5748 0.0003504 ***
c4     -0.261420  0.105690 -2.4735 0.0133809 *
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

dF/dx is for discrete change for the following variables:

[1] "c4"
```

### 4.5.3 표준화된 해석

예측변수가 여러 개 있는 경우, 예측변수의 효과를 비교하기 위해서  $\{\hat{\beta}_j\}$ 의 크기를 비교하고자 한다. 이항 예측변수의 경우, 다른 예측변수들이 주어졌을 때, 조건부로 그 오즈비를 비교하는 것과 같다. 양적 예측변수의 경우, 만약 예측변수들이 모두 동일한 단위를 갖고 있어서 한 단위의 변화가 각각의 예측변수에서 동일한 의미를 가질 경우에는 이러한 비교가 의미가 있다. 그러나 그 외의 경우에는 비교하는 것이 아무런 의미가 없다.

서로 다른 단위를 가지는 양적 예측변수들의 효과를 비교하기 위한 대안으로 **표준화 계수**를 사용한다. 모형은 표준화된 예측변수를 적합하는데 예측변수  $x_j$  대신에  $(x_j - \bar{x}_j)/s_{x_j}$  으로 대체하여 적합한다.

$s_{x_j}$  는  $x_j$  의 표본표준오차를 의미한다. 따라서  $\hat{\beta}_j$  는 다른 변수들이 고정되었을 때 예측변수  $x_j$  의 표준오차만큼의 변화에 대한 효과를 나타낸다. 예측변수  $x_j$  의 표준화된 추정값은 표준화되지 않은 추정값  $\hat{\beta}_j$  에  $s_{x_j}$  를 곱한 것과 같다.

예로, 암참게 자료에 대해서 너비와 색깔(점수 1,2,3,4)을 양적 설명변수로 포함하는 모형의 예측식은

$$\text{logit}(\hat{\pi}) = -10.071 + 0.458x - 0.509c_4$$

이때 설명변수들의 분산값들에 차이가 있기 때문에 이와 같은 예측식으로부터 각 설명변수들의 효과가 비슷하다고 결론을 내리면 잘못된 것이다.

너비의 경우  $\bar{x} = 26.30$  이고  $s_x = 2.11$  인 반면, 색깔은  $\bar{c} = 2.44$  이고  $s_c = 0.80$  이다.

설명변수에 대한 효과의 추정값은  $(2.11)(0.458) = 0.97$  이고,  $(0.80)(-0.509) = -0.41$  이다.

다른 변수를 고정시켰을 때 너비의 표준오차 증가 효과는 색깔의 표준오차 증가 효과의 두 배 이상일 것으로 추정된다

## 4.6 예측력 요약 : 분류표, ROC 곡선, 다중상관성

모형을 비교하기 위해 각 모형을 적합시켜 반응변수의 결과값을 얼마나 잘 예측할 수 있는지 나타내는 예측력을 서로 비교해 보는 것이 매우 유용하다. 이 절에서는 예측력을 평가하는 세 방법을 제시한다.

### 4.6 예측력 요약 : 분류표

분류표는 이항결과변수  $y$ 와 예측된 결과를 분류한 표이다. 주어진 분계점 값  $\pi_0$ 을 기준으로  $\hat{\pi}_i > \pi_0$  일 경우 관측값  $i$ 의 예측값은  $\hat{y} = 1$  이 되고,  $\hat{\pi}_i \leq \pi_0$  일 경우  $\hat{y} = 0$  이 된다.

가능한  $\pi_0$ 의 값으로  $\pi_0 = 0.50$  을 쓸 수 있다. 그러나 만약 간측값들 중에서  $y=1$ 인 비율이 낮은 경우에는 모형의 적합값은 절대로  $\hat{\pi}_i > 0.50$  이 될수 없으므로  $\hat{y}=1$  로 예측할 수 없다.

이와 반대로 관측값들이  $y=1$ 인 비율이 높을 경우, 모형의 적합값은 항상  $\hat{\pi}_i > 0.50$  이 될 수 있으므로 항상  $\hat{y}=1$  로 예측한다. 다른  $\pi_0$  의 이항결과변수가 1인 표본비율값으로 설정할 수 있다.

이 값은 상수항만 있는 모형의  $\hat{\pi}_i$  값이다.

너비와 색갈 요인을 설명변수로 갖는 암참개가 부수체를 가질 확률에 대한 모형을 예로 들며 173마리의 암참개 중 111마리가 부수체를 가지며 표본 비율은 0.6416이다.

```
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat", head$  
> prop <- sum(Crabs$y)/nrow(Crabs)  
> prop  
[1] 0.6416185  
>  
> fit <- glm(y~width + factor(color), family=binomial, data=Crabs)  
> predicted <- as.numeric(fitted(fit) > prop)  
> xtabs(~ Crabs$y + predicted)  
predicted  
Crabs$y  0  1  
      0 43 19  
      1 36 75
```

## 4.6 예측력 요약: 분류표, ROC 곡선, 다중상관성

$y=0$ 인 62개의 표본 중 모형에서  $\hat{y}=0$  으로 예측된 표본의 수는 43개이다. 111개의  $y=1$ 인 표본 중 모형에서  $\hat{y}=1$  로 예측된 표본의 수는 75개이다.

▶ 표 4.4 참게 자료에서 너비와 색깔에 대한 지시변수를 예측변수로 고려한 경우의 분류표

실제	예측, $\pi_0 = 0.6416$		예측, $\pi_0 = 0.50$		합계
	$\hat{y}=1$	$\hat{y}=0$	$\hat{y}=1$	$\hat{y}=0$	
$y=1$	75	36	96	15	111
$y=0$	19	43	31	31	62

위 표는  $\pi_0 = 0.6416$  이고  $\pi_0 = 0.50$  인 경우 예측값에 대한 분류표이다.  
예측력을 평가할 수 있는 유용한 두 가지 요약값은 다음과 같다.

$$\text{민감도} = P(\hat{y}=1|y=1), \text{특이도} = P(\hat{y}=0|y=0)$$

## 4.6 예측력 요약 : 분류표, ROC 곡선, 다중상관성

2.1.2절에 의학 진단 검사에서 예측력을 평가하기 위한 측도들을 소개하였다.

$\pi_0 = 0.6416$ 일때, <표 4.4>에서 민감도의 추정값은  $75/111=0.676$ 이고 특이도 추정값은  $43/62=0.694$ 이다.

전체 자료에서 맞게 분류된 표본의 비율은  $(75+43)/(111+62)=0.682$ 이다.

이 추정값은 다음과 같이 민감도와 특이도의 가중평균값이다.

$$\begin{aligned} P(\text{올바른 분류}) &= P(y=1| \hat{\pi}=1) + P(y=0| \hat{\pi}=0) \\ &= P(\hat{\pi}=1|y=1) + P(\hat{\pi}=0|y=0)P(y=0) \\ &= \text{민감도}[P(y=1)] + \text{특이도}[1 - P(y=1)] \end{aligned}$$

예측력을 평가하기 위한 위와 같은 요약값은  $y_i$  를 포함하는 자료를 적합시켜  $\hat{\pi}_i$  를 추정하기 때문에 실제보다 예측력을 더 높게 평가할 수 있다. 따라서  $y_i$  를 제외한 나머지  $n-1$ 개의 관측값들로부터  $\hat{\pi}_i$  를 추정하는 러브-원-아웃 교차검증법(LOOCV)를 사용함으로써 모형의 예측력을 평가하는 것이 더 좋다.

암참게 자료에 대해서 LOOCV 방법을 실시하면 올바르게 분류된 표본의 비율은 0.671이다.

색깔만을 예측변수로 삼을 경우 이 비율은 0.642, 너비만을 예측변수로 삼을 경우 0.659,

너비와 색깔에 대한 지시변수를 예측변수로 삼을 경우 0.682이다.

하지만 이 분류표는 연속형 예측값  $\hat{\pi}$  를 이항값으로 변환시켜야 하고  $\pi_0$ 의 값을 임의로 정해야 하며  $y=1$  또는  $y=0$ 인 표본들의 상대적인 크기가 결과값에 민감하게 영향을 준다는 한계를 가지고 있다.

## 4.6.2 예측력 요약: ROC 곡선

수신기 작동 특성(receiver operating characteristic, ROC) 곡선은 가능한 모든 기준값  $\pi_0$ 에 대하여 예측의 민감도와 특이도를 보여주는 그림이다. ROC 곡선은 가능한 모든  $\pi_0$  값에 대한 예측력을 요약하는 값이기 때문에 분류표보다 더 많은 정보를 가지고 있다.

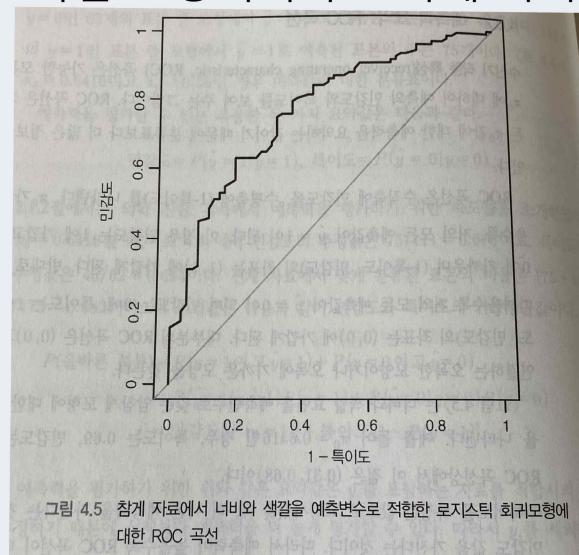
ROC 곡선은 수직축에 민감도를, 수평축에 (1-특이도)를 나타낸다.

$\pi_0$  가 0에 가까울수록, 거의 모든 예측값이  $\hat{y} = 1$  이 된다. 이 경우 민감도는 1에 가깝고 특이도는 0에 가까우며 (1-특이도, 민감도)의 좌표는 (1,1)에 가깝게 된다.

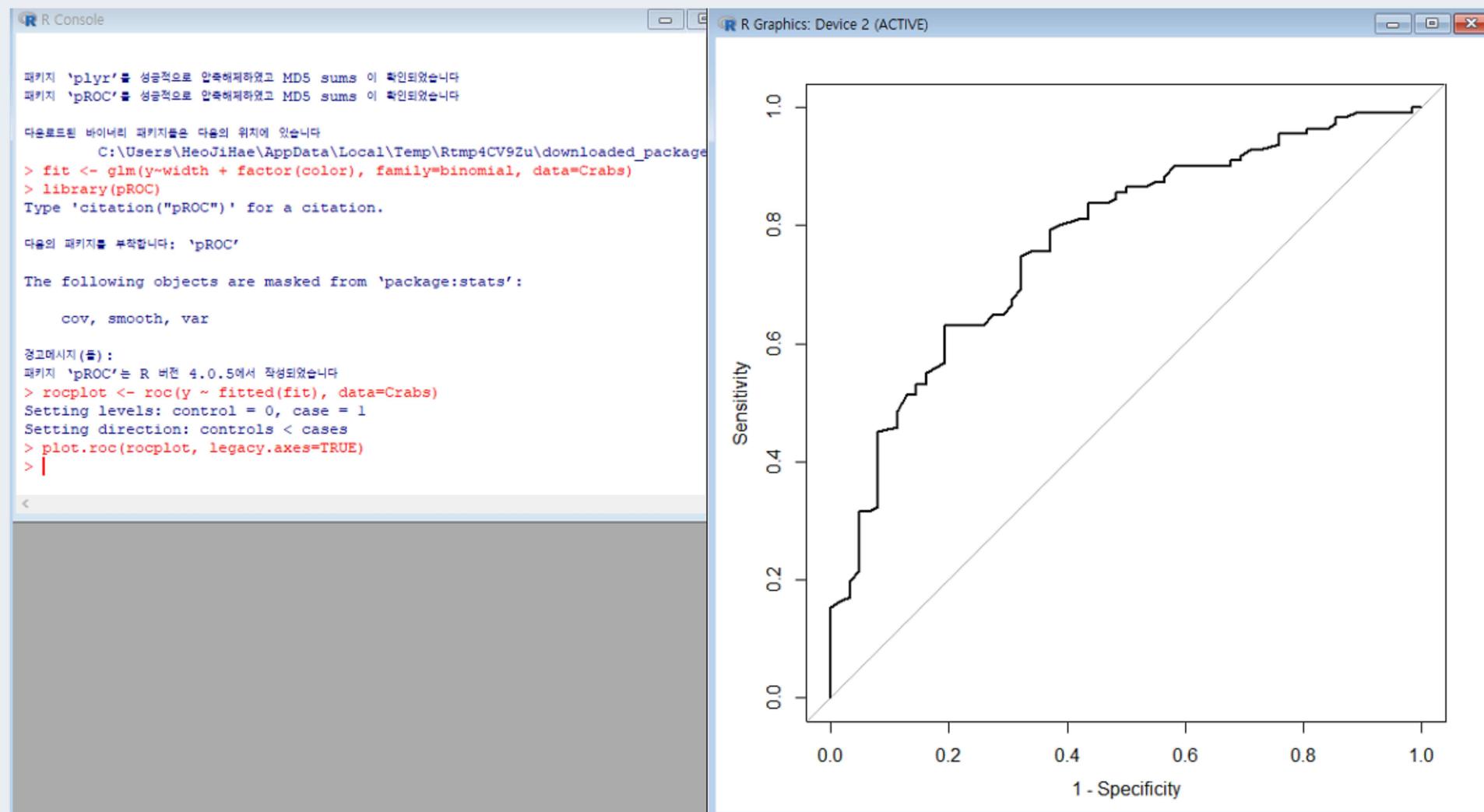
반대로  $\pi_0$ 가 1에 가까울수록 거의 모든 예측값이  $\hat{y} = 0$  이 되며 민감도는 0에, 특이도는 1에, (1-특이도, 민감도)의 좌표는 (0,0)에 가깝게 된다.

대부분의 ROC곡선은 (0,0)과 (1,1)을 연결하는 오목한 모양이거나 오목에 가까운 모양을 갖는다.

그림은 너비와 색깔 요인을 예측변수로 갖는 암참개 모형에 대한 ROC 곡선을 나타낸다.



## 4.6.2 예측력 요약: ROC 곡선



## 4.6.2 예측력 요약: ROC 곡선

특이도를 특정한 값으로 고정시켰을 때, 더 나은 예측력을 가진다는 것은 더 높은 민감도 값을 가진다는 것이다. 따라서 예측력이 높을수록 ROC곡선은 더 높게 그려진다.

따라서 ROC곡선 아래의 넓이는 예측력을 요약하는 하나의 값을 나타나게 된다.

ROC곡선의 넓이가 넓을수록 예측력이 높다. 예측력에 대한 측정치를 일차성 지수라고 부른다.

$y_i = 1$  이고  $y_i = 0$  인 관측치  $(i,j)$ 의 모든 대응쌍을 생각해보자.

일차성 지수는 예측값과 관측값이 일치할 확률을 추정하며 이것은  $y$ 값이 큰 관측값이 값 또한 큰 값을 가진다는 것을 의미한다.

일차성 지수가 0.50이라는 것은 임의로 예측한 것보다도 예측이 더 잘되지 않았다는 것을 의미한다.

이는 상수항만을 가지는 모형에 대응되며 ROC곡선은  $(0,0)$ 과  $(1,1)$ 을 연결하는 직선이 된다.

암참게 자료에서 모형별로 구한 표본 일차성 지수는 다음과 같다.

색깔 요인 예측변수만 포함 -> 0.639

너비만 예측변수로 포함 -> 0.742

너비와 색깔 요인을 같이 포함 -> 0.771

너비와 어두운 색깔을 나타내는 지시변수 -> 0.772

### 4.6.3 예측력 요약 : 다중상관성

GLM에서 예측력을 평가하는 또 다른 요약값은 관찰된 반응값  $\{y_i\}$ 과 모형에서의 적합값  $\{\hat{\mu}_i\}$  간에 구한 상관계수 R이다. 선형모형을 적합시키는 최소제곱법에서 R은 반응변수와 설명변수들 간의 **다중상관계수**이다.  $R^2$ 은 예측변수들에 의해 설명되는  $y$  변동의 비율을 나타낸다.

$R^2$ 과 비교했을 때 R은 기존 자료의 척도를 사용하고 효과의 크기에 근사적으로 비례하는 값을 가진다는 장점을 가지고 있다.

이항회귀모형에서 R은 n개의 이항관측값  $\{y_i\}$ (0또는 1) 와 적합된 비율  $\{\hat{\pi}_i\}$  간의 상관성을 나타낸다.  $y$ 의 이산적인 성질로 인해, 특히 0과 1 값들의 빈도수 불균형이 클수록 가능한 R값들의 범위가 제한될 수 있다.

다른 상관성 측도들도 마찬가지로 R값은 설명변수들이 관측된 영역에 의해서 영향을 받는다. 그럼에도 R은 같은 자료에 대한 적합한 여러 다른 모형들의 적합도를 비교할 때 유용하다.

#### 4.6.3 예측력 요약 : 다중상관성

```
> fit <- glm(y~width + factor(color), family=binomial, data=Crabs)
> cor(Crabs$y, fitted(fit))
[1] 0.4522131
```

너비와 어두운 색깔을 가지는 여부에 대한 지시변수를 사용하는 더 간단한 모형 또한 자료를 더 잘 적합시키며 이때의 상관 계수는 R=0.447이다.  
너비만을 설명변수로 사용할 경우 R=0.402이다.

보통의 최소제곱회귀모형에서와 같이 이 측정값의 제곱값을 분산을 설명하는 비율로 해석할 수는 없다. 이항자료에 대하여 이와 비슷한 해석을 하기 위해 다양한 측도들이 제시되었다.