3-2-1 분류 분석과 로지스틱 회귀모형

분류 분석은 반응 변수가 알려진 다변량 자료를 이용하여 모형을 구축하고, 이를 통해 새로운 자료에 대한 예측 및 분류를 수행하는 것이 목적

- 반응 변수가 범주형인 경우에 예측 모형은 새로운 자료에 대한 분류가 주목적
- 반응 변수가 연속형인 경우에는 그 값을 예측하는 것을 주목적
- 따라서 예측과 분류는 유사한 의미로 사용
- 예측 및 분류 기법은 목표 마케팅 성과 예측 의학 진단 사기 검출 제조 등 다양한 분야에 이용

로지스틱 회귀모형의 이해

로지스틱 회귀(logistic regression) 모형은 반응변수가 범주형인 경우(0 or 1) 적용하는 회귀분석 모형

로지스틱 회귀모형은 설명변수의 값이 주어질 때,

- 특정 종속변수 집단에 속할 확률을 추정하여 특정 임계값을 설정하여 분류작업으로 진행

이 때 모형 적합을 통해 추정된 확률은

- "사후확률(posterior probability)"이라고 부르기도 함

기본적인 다중 로지스틱 회귀모형의 수식은 아래와 같음

$$log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$$

$$\pi(x) = P(Y = 1|x), x = (x_1, ..., x_k)$$

위의 식은 승산비(odds)에 로그를 취한 식

- 그렇기에 해석에 있어서 단순 확률이라고 읽으면 안된다.
- 승산비란 [성공확률(주류) = p] / [실패확률(비주류) = (1-p)]이다.

$$\frac{\pi(x)}{1-\pi(x)}$$

결론

- 로지스틱 회귀는 “확률이 아니라 **로그 오즈를 선형적으로 모델링한다**”는 의미

정리

오즈(odds) : 성공할 확률 대 실패할 확률의 비율, 승산비

$$\frac{\pi(x)}{1-\pi(x)}$$

- 분자인 pi(x)는 "성공 확률" (예: 질병이 있을 확률, 비만일 확률, 합격할 확률 등)
- 문모인 1 - pi(x)는 "실패 확률" (그렇지 않을 확률)
- pi(x)/(1 - pi(x)): 식을 **오즈(odds)하고 부름**

  → 성공이 실패보다 몇 배 더/덜 일어날 것 같은가?를 수치로 표현

항목	설명	예시 (확률이 0.8일 때)
성공할 확률	pi(x) = P(Y = 1 | x)	0.8
실패 확률	1 - pi(x)	0.2
오즈 (odds)	pi(x) / (1 - pi(x))	0.8 / 0.2 = 4

🎰 도박 (gambling)

- "The odds of winning this slot machine are 1 in 1000."

> (이 슬롯머신에서 이길 가능성은 1,000분의 1이다.)

- "The odds are 2 to 1"

  → 성공할 가능성이 실패할 가능성보다 2배 많다는 의미

오즈(odds) 값을 log(odds)를 취하면 바로 **로그 오즈(log odds)** 또는 **로짓(logit) 함수**가 됨:

$$\mathrm{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

확률예측 수식

$$P(Y=1 \mid \mathbf{x}) = \frac{1}{1+\exp\left(-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)\right)}$$

📘 설명

- x = (x1, x2, ..., xk): 입력 변수 벡터
- β0: 절편 (intercept)
- β1 ~ βk: 각 입력 변수 xi에 곱해지는 계수 (회귀계수)
- z = β0 + β1·x1 + β2·x2 + ... + βk·xk: 선형 결합
- P(Y = 1 | x) = 1 / (1 + exp(−z)): 시그모이드 함수로 계산한 확률값
- 이 확률은 Y가 1일 가능성을 의미함 (예: '비만일 확률', '합격할 확률' 등)
- 1 / (1 + exp(−z))는 바로 **시그모이드 함수 σ(z)**임

그러므로 다음과 같이 시그모이드함수로 표현

$$P(Y=1|\mathbf{x})=\sigma(\beta_0+\beta_1x_1+\cdots+\beta_kx_k)$$

📘 단계별 유도 과정

1단계: 확률 표현

로지스틱 회귀분석의 수식은 다음과 같다:

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p)}}$$

위와 같은 표현

$$P(Y=1\mid\mathbf{x})=\frac{1}{1+\exp\left(-(\beta_0+\beta_1x_1+\cdots+\beta_kx_k)\right)}$$

분모, 분자를 위치를 바꾸면

양변을 뒤집기 위한 준비를 해보자:

$$\Rightarrow 1+\exp(-z)=\frac{1}{P}$$

$$\Rightarrow \exp(-z)=\frac{1-P}{P}$$

양변에 로그(log)를 취하면:

$$-z=\log\left(\frac{1-P}{P}\right)\Rightarrow\log\left(\frac{P}{1-P}\right)=z$$

2단계: z를 선형결합으로 치환

또는 로짓(logit) 함수 형태기 됨

$$\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right) = \beta 0 + \beta 1 \cdot x1 + \ldots + \beta k \cdot xk$$

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

여기서:

- P(Y = 1 | X)는 주어진 X 값에서 Y가 1일 확률
- β_0는 절편
- β_1, β_2, ..., β_p는 각 변수의 회귀계수
- X_1, X_2, ..., X_p는 독립변수(설명변수)

이게 바로 로그 오즈 수식이야!

- 즉, 시그모이드 함수의 역함수 형태가 바로 logit 함수

시그모이드 함수는 오른쪽으로 표현: 왼쪽의 분모, 분자에 ez를 모두 곱하면 오른쪽이 됨

$$\frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$$

다음 결과

$$\frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

그러므로 다음 결과

$$\pi(x) = \frac{exp(\alpha + \beta_1 x_1 + ... + \beta_k x_k)}{1 + exp(\alpha + \beta_1 x_1 + ... + \beta_k x_k)}$$

$$\pi(x) = \frac{1}{1 + exp\{-(\alpha + \beta_1 x_1 + ... + \beta_k x_k)\}}$$

실습

```
data(iris)
a <- subset(iris, Species == "setosa" | Species == "versicolor")
a$Species <- factor(a$Species)
# 함수 glm은 Generalized Linear Model
b <- glm(Species ~ Sepal.Length, data = a, family = binomial)
summary(b)

Call:
glm(formula = Species ~ Sepal.Length, family = binomial, data = a)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -27.831      5.434  -5.122 3.02e-07 ***
```

```
Sepal.Length    5.140      1.007   5.107 3.28e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.629  on 99  degrees of freedom
Residual deviance:  64.211  on 98  degrees of freedom
AIC: 68.211

Number of Fisher Scoring iterations: 6
```

요약 먼저 말하자면 이건 **로지스틱 회귀모형으로 붓꽃의 품종(setosa vs versicolor)을 꽃받침 길이(Sepal.Length)만으로 구분한 결과**야. 모델이 통계적으로 유의하며, 꽃받침 길이가 길어질수록 versicolor일 확률이 커진다는 걸 보여주고 있어.

🌸 R glm() 로지스틱 회귀 분석 결과 해석

📌 1. 분석 배경

- `iris` 데이터에서 **setosa와 versicolor** 두 종만 추출해 이진 분류(binary classification)로 구성함
- 목표: `Sepal.Length` (꽃받침 길이)로 품종(`Species`)을 예측
- 사용 모델: **로지스틱 회귀 (`glm(family = binomial)`)**

📌 2. 회귀식 구조

- 모델식:log(P(setosa ∣ x)/P(versicolor ∣ x))=−27.831+5.140×Sepal.Length

$$\log\left(\frac{P(\text{versicolor} \mid x)}{P(\text{setosa} \mid x)}\right) = -27.831 + 5.140 \times \text{Sepal.Length}$$

여기서:

- 종속변수 `Species` 는 범주형 → 내부적으로 `"setosa"` 를 0, `"versicolor"` 를 1로 변환
- 따라서 **π(x) = P(Species = versicolor | Sepal.Length)** 를 예측하는 모델임

📌 3. 계수 해석 (summary 출력 기준)

계수	추정값	의미
(Intercept)	-27.831	꽃받침 길이가 0일 때 로그 오즈 값
Sepal.Length	+5.140	꽃받침 길이 1cm 증가 시, 로그 오즈 +5.14 상승

- z-value와 p-value가 모두 ***로 유의 수준 0.001 이하임 → **계수들이 통계적으로 유의미**함
- 꽃받침 길이가 길수록 **versicolor일 확률이 급격히 증가**함

📌 4. 모델 적합도

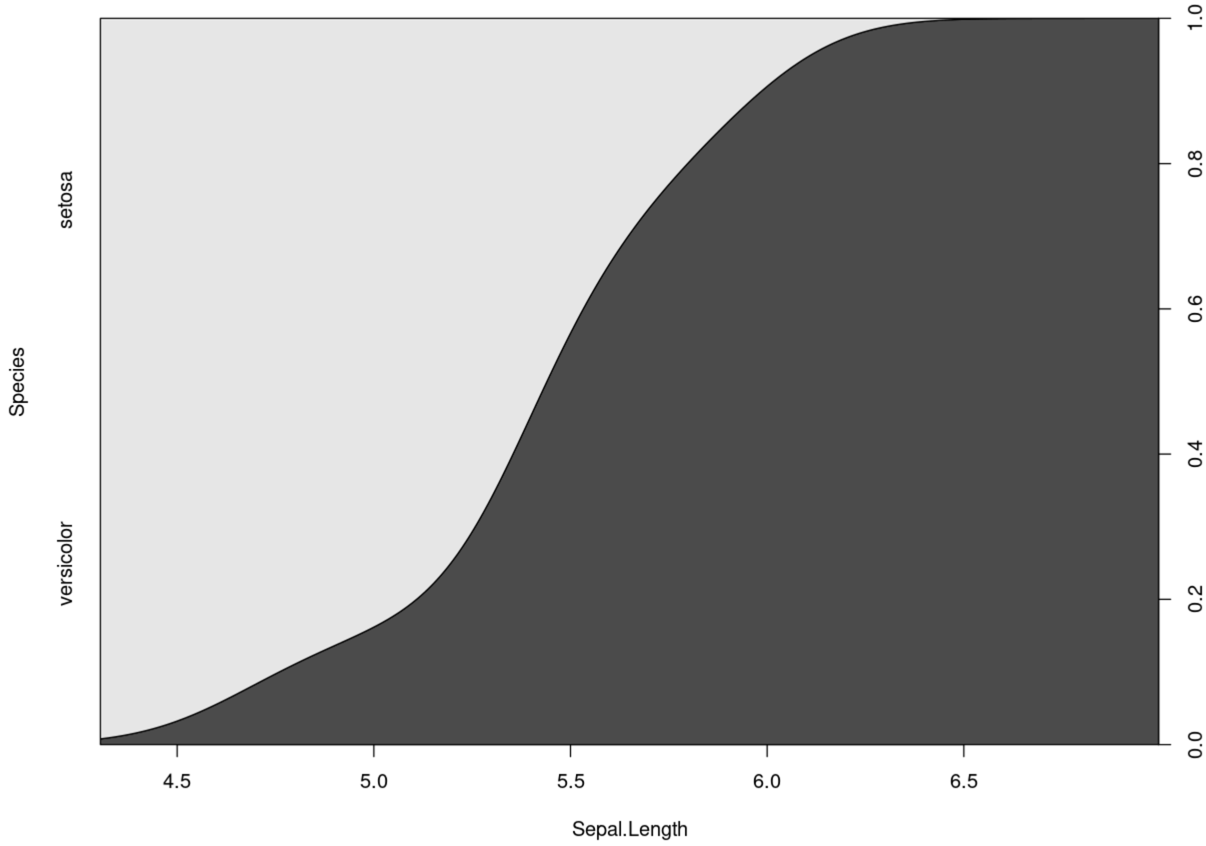
항목	값	해석
Null deviance	138.629	독립변수 없이 모델링한 deviance
Residual deviance	64.211	Sepal.Length 포함한 deviance (낮을수록 좋음)
AIC	68.211	모델의 정보 기준 (낮을수록 좋은 모델)

- **잔차 deviance가 크게 감소** → Sepal.Length가 모델에 큰 기여를 했다는 의미
- **AIC 68.211** → 단일 변수로는 상당히 좋은 분류 모델

📌 5. 결론 요약

- 이 로지스틱 회귀모델은 꽃받침 길이만으로도 setosa와 versicolor를 **정확히 구분** 가능
- 길이가 길수록 setosa가 아닌 **versicolor일 가능성이 커짐**
- 모델의 **설명력도 높고 통계적으로 매우 유의함**

```
# cdplot(): 함수로 조건부 분포 시각화(Conditional Density Plots)
cdplot(Species ~ Sepal.Length, data=a)
```



이 그래프는 `cdplot()` 함수로 만든 **조건부 밀도 도표(Conditional Density Plot)**로,

`Sepal.Length`가 변할 때 `Species`가 **versicolor**일 확률이 어떻게 달라지는지를 시각적으로 보여줘.

🎯 cdplot(Species ~ Sepal.Length) 결과 해석

✔️ 1. x축: Sepal.Length (꽃받침 길이)

- 4.3cm ~ 7.0cm 사이의 연속형 변수

✔️ 2. y축: 조건부 확률 (0~1 범위)

- 이건 "해당 꽃받침 길이일 때 **품종이 무엇일 확률인가?**"를 의미해
- 위쪽: setosa일 확률

- 아래쪽: versicolor일 확률

> y축이 "Species"로 보이지만 실제로는 확률의 누적 영역으로 시각화된 거야.

✔️ 3. 색 영역

- **밝은 회색 (위)**: setosa일 확률
- **짙은 회색 (아래)**: versicolor일 확률
- 각 세로선에서 회색 높이의 비율이 바로 `P(Species | Sepal.Length)` 값을 의미함

📊 핵심 패턴 설명

Sepal.Length 구간	우세한 품종	의미 요약
4.3 ~ 5.0	setosa	setosa일 확률이 거의 1 (분류 명확)
5.0 ~ 6.0	점점 변화	두 품종이 섞여 있고 확률 곡선이 변곡됨 (판단 애매한 구간)
6.0 이상	versicolor	versicolor일 확률이 거의 1 (분류 명확)

- **5.4cm 근처**가 분류 경계처럼 보임 (S자 형태의 중간 지점)
- 이는 로지스틱 회귀의 시그모이드 함수 중심과도 일치함

종료