Logistic Regression
(로지스틱 회귀)

Prof. Hyerim Bae
Department of Industrial Engineering, Pusan National University
hrbae@pusan.ac.kr

예측과 분류의 차이

- 예측(Prediction)이란?
  - 데이터의 연속형 값을 알아맞히는 것
    - 연속형 값의 예) 사람의 키, 자동차의 몸무게 등

- 분류(Classification)란?
  - 데이터의 범주형 값을 알아맞히는 것
    - 범주형 값의 예) 대출 승인 여부, 자동차의 품질 등

로지스틱 회귀

- 선형 회귀 분석의 개념에서, 종속 변수가 범주형인 상황으로 확장

- 특히 설명하거나, 분류해야 할 경우 널리 사용됨
  - 고객을 반납/비반납 고객으로 분류 (분류 문제)
  - 남자 최고경영진과 여자 최고경영진으로 구별하는 요인 찾기 (프로파일링)

- 이진 분류
  - 예) Y=0 or Y=1

- 2 Steps
  - 각 클래스에 속하는 확률을 추정
  - 각 관측치를 이들 클래스 중 하나로 분류하기 위해 확률값에 대한 분류 기준값을 적용

로짓(Logit)

- 목표: 입력변수가 존재하며, 결과값이 0 혹은 1로 구성된 함수를 찾는 것
- 선형 회귀분석의 결과값 Y 대신 로짓이라고 하는 함수를 사용
- 로짓은 입력변수의 선형 함수로 모델링 할 수 있음
- 로짓은 확률로 다시 매핑될 수 있으며, 확률은 클래스로(0 또는 1) 매핑할 수 있음

1단계: 로지스틱 함수(Logistic Response Function)

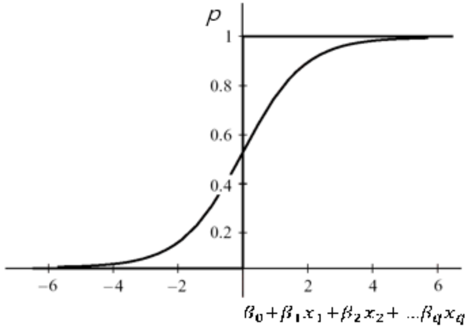
- p = 클래스 1에 속할 확률
- $0 \leq p \leq 1$을 만족하는 함수로, p를 예측변수와 연관시켜야 함
- 표준화된 선형 함수(아래 그림 참조)

Want to guarantee that Y exists in [0, 1]

$$p_{LR} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_q x_q$$

q = number of predictors

로지스틱 함수



$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_q x_q)}}$$

2단계: 오즈(The Odds)

- 오즈(The Odds)의 정의
  - 클래스 1에 속할 오즈 정의

p = 사건이 발생할 확률

$$Odds = \frac{p}{1-p}$$

"클래스 0에 속할 확률에 대한 클래스 1에 속할 확률"

$$p = \frac{Odds}{1+Odds}$$

오즈와 입력 변수

$$Odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q} \quad \cdots$$

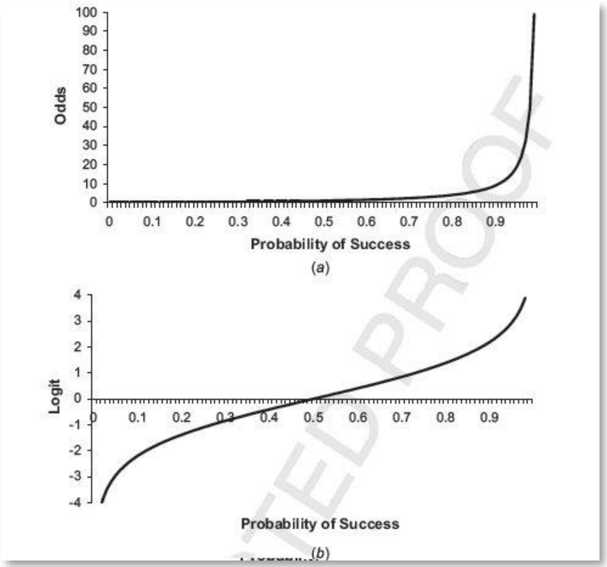
x_j가 한단위 증가하면?

3단계: 양변을 로그 형태로 변환

$$\log(Odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

로짓

- 즉, 로짓은 입력 변수의 선형 함수임
- $-\infty$에서 ∞의 값을 가질 수 있음
- 로짓, 오즈, 확률간의 관계



개인 대출 서비스

- 출력 변수: 은행 대출을 받는다(0/1).
- 입력 변수: 연령, 소득, 담보 대출건, 증권계좌 등

데이터 전처리

- 60% 비중: 학습용으로, 40% 비중: 검증용으로 분할
- 범주형 입력 변수들은 0 혹은 1로 표현되는 더미 변수 생성

$$EducProf = \begin{cases} 1 \text{ if education is } \textit{Professional} \\ 0 \text{ otherwise} \end{cases}$$

$$EducGrad = \begin{cases} 1 \text{ if education is at } \textit{Graduate} \text{ level} \\ 0 \text{ otherwise} \end{cases}$$

$$Securities = \begin{cases} 1 \text{ if customer has securities account in bank} \\ 0 \text{ otherwise} \end{cases}$$

$$CD = \begin{cases} 1 \text{ if customer has CD account in bank} \\ 0 \text{ otherwise} \end{cases}$$

$$Online = \begin{cases} 1 \text{ if customer uses online banking} \\ 0 \text{ otherwise} \end{cases}$$

$$CreditCard = \begin{cases} 1 \text{ if customer holds Universal Bank credit card} \\ 0 \text{ otherwise} \end{cases}$$

단일 예측변수 모델

- 소득에 대한 대출의 수용을 모델링

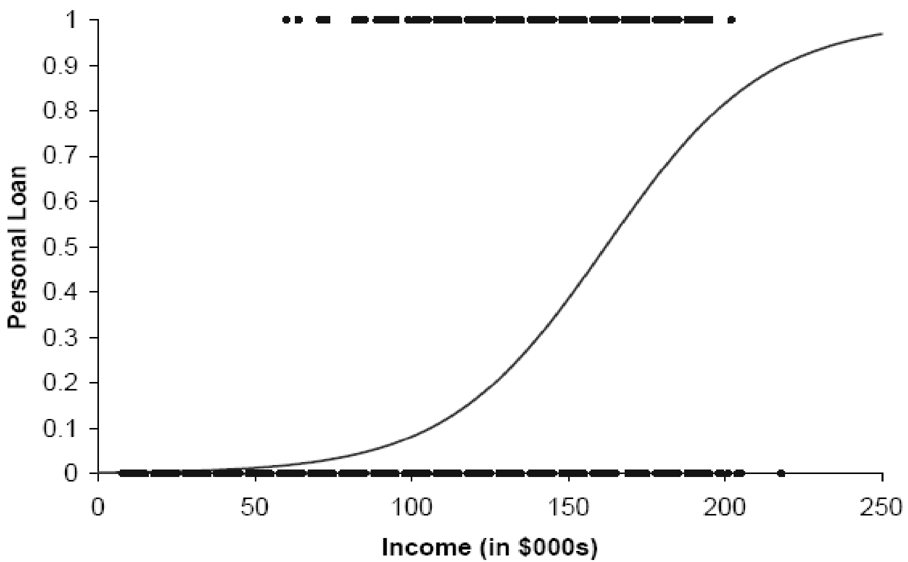
$$\text{Prob}(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

- 계수 추정: β_0 = −6.3525, β_1 = 0.0392

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{6.3525 - 0.0392x}}$$

관계 시각화

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{6.3525 - 0.0392x}}$$



마지막 단계: 분류

- 모델은 1이라고 추정할 확률을 도출
  - Cutoff level을 설정하여 분류로 변환
  - 만약, 추정한 확률 〉 Cutoff라면, 1으로 분류

Cutoff를 결정하는 방법

- 보편적인 초기 Cutoff 선택 기준: 0.5

- 추가 고려 사항
  - 분류 정확도 최대화
  - Sensitivity 최대화(TP/(TP+FN))
  - FP(False Positive) 최소화: 참이라고 예측했을 때, 실제 결과는 거짓인 경우
  - 분류가 잘못되었을 때, 예상되는 비용의 최소화(비용 지정 필요)

예제

- β의 추정치는 최대 우도 추정법이라는 반복 과정을 통해 도출됨
  - 주어진 데이터를 얻을 가능성을 최대로 하는 추정치를 찾는 방법
  - 컴퓨터 프로그래밍을 사용하여 반복 및 추정

- 예측 모형: 12개 입력 변수를 포함해 보자.

- XLMiner의 출력은 로짓에 대한 계수와 개별 항의 오즈를 제공함

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

Figure 10.3: Logistic regression coefficient table for personal loan acceptance as a function of 12 predictors.

로짓 추정 방정식

$$\begin{aligned}\text{logit} = &-13.201 - 0.045 Age + 0.057 Experience + 0.066 Income + 0.572 Family \\ &+0.18724874 CCAvg + 0.002 Mortgage - 0.855 Securities + 3.469 CD \\ &-0.844 Online - 0.964 Credit\ \ Card + 4.589 EducGrad + 4.523 EducProf\end{aligned}$$

오즈 방정식

음의 계수 양의 계수

$$\text{odds}(Personal\ Loan = \text{Yes}) = e^{-13.201}(0.956)^{Age}\ (1.058)^{Experience}\ (1.068)^{Income}$$
$$\cdot\ (1.771)^{Family}\ (1.206)^{CCAvg}\ (1.002)^{Mortgage}$$
$$\cdot\ (0.425)^{Securities}\ (32.105)^{CD}\ (0.430)^{Online}$$
$$\cdot\ (0.381)^{CreditCard}(98.405)^{EducGrad}\ (92.086)^{EducProf}$$

확률로 변환

$$p = \frac{Odds}{1 + Odds}$$

오즈 및 확률 해석

- 분류의 경우, 일반적으로 Cutoff 값을 고려하여 확률 값을 사용함

- 설명을 목적으로 하는 오즈는 다음과 같이 해석함:
  - $x_2, x_3, \ldots, x_q$를 일정하게 유지하면서, x_1을 한 단위 증가시킨다면, b_1은 클래스 1에 속하는 오즈가 증가하는 요인임
  - 예측 변수가 가변수일때,
    - CD에 대한 $odds = 32.015 \rightarrow CD$ 계좌를 가지지 않는 고객에 비하여 대출 제안을 수락할 $odds$
  - 가변수가 아닌 연속형 변수일 때,
    - 예: x_j가 3 → 4로 증가할 때와 30 → 31로 증가할 때, p에 미치는 효과가 상이함
  - 오즈 비
    - 두 개의 범주사이의 오즈 값의 비
    - 예) 전문 교육과 대학원 교육을 받은 고객에 대한 대출 제안 수락 여부: 오즈 비가 1보다 크면, 전문 교육을 받은 고객이 대학원 교육을 받은 고객보다 대출 수락할 확률이 높다

정리

- 로지스틱 회귀 분석은 범주형 변수와 함께 사용된다는 점을 제외하고는 선형 회귀분석과 유사함
- 설명하려는 문제(=프로파일링), 예측하려는 문제(=분류)에서 사용할 수 있음
- 예측 변수는 로짓이라는 비선형 함수를 통해 변환시킬 수 있음
- 선형 회귀분석과 마찬가지로 변수 선택을 통해 예측 변수를 줄일 수 있음
- 로지스틱 회귀 분석을 세 개 이상의 클래스로 일반화 할 수 있음
  - 기존: 두 개의 클래스(XLMiner에는 없음)

부록

3.3.1 선형성(linearity)

진단방법 ①(설명변수와 종속변수) 산점도 → 이차 함수 형태

②잔차와 예측치 산점도 → 이차 함수 형태

해결방법 ①설명 변수의 이차항이나 다차항을 삽입한다.

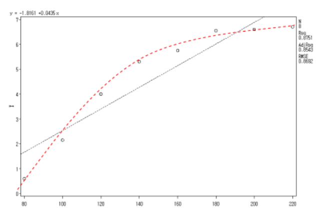
산점도를 보면 종속변수와 설명변수의 직선(선형) 관계를 진단할 수 있다. 잔차와 예측치의 산점도가 일정한 함수 형태를 가지면(일반적으로 이차 함수) 선형성이 무너지게 되는데 이를 해결하려면 설명변수의 이차항을 설명변수로 추가한다. 이차항을 추가할 때는 설명변수를 표준화 한 후 넣으면 다중공선성 문제가 완화된다. (다음 페이지 참고)

EXAMPLE 3-2 선형성 파괴: 이차 관계

```
data quard;
    input y x @@;
    cards;
0.6 80 6.7 220 5.3 140 4 120
6.55 180 2.15 100 6.6 200 5.75 160
run;

proc reg data=quard;
    model y=x;
    plot y*x;
    plot student.*predicted.;
run;
```

종속변수와 설명변수의 산점도를 보면 직선 관계라고 보기 어렵다. 이차 함수 관계에 가깝다.



3.3.2 등분산성(homoscedasticity)

진단방법 ①잔차와 예측치 산점도, 나팔 모양

해결방법 ①가중최소자승법, WLS(Weighted Least Square) 사용한다.

②종속변수변환. 일반적으로 LOG 변환을 하는 것이 일반적이다.

잔차와 예측치 산점도에서 나팔 모양이면 오차의 분산이 예측치가 커짐에 따라 커지거나 작아지고 있음을 의미하므로 등분산 가정이 무너지게 된다. 이런 경우 가중최소자승 추정치를 이용하거나 종속변수변환을 실시한다. 등분산의 경우 일반적으로 오차의 분산은 $V(e_i)=\sigma_i^2=\sigma^2/w_i$ 으로 가정되고 가중최소자승가중치로 $w_i=1/y_i^2$, 혹은 $w_i=1/x_i^2$ 을 주로 사용한다.

WLS(Weighted Least Square)

$\min_{\alpha,\beta}\sum w_i(y_i-\alpha-\beta x_i)^2$ 인 $\hat{\alpha},\hat{\beta}$ 을 WLS 추정치라 한다. 일반적으로 가중치 w_i 는 $1/\sigma_i^2$ (σ_i^2

을 알고 있을 때, 그러나 실제 알지 못한다) 혹은 $1/x_i^2$, $1/\hat{y}_i^2$ 등을 사용한다. 단순회귀의 잔차분석은 잔차와 예측치 산점도에 주로 의존하므로 $1/\hat{y}_i^2$ 을 주로 사용한다. 다중회귀에서는 문제가 되는 설명변수를 이용한 가중치 $1/x_i^2$ 을 사용하기도 하지만 판단이 쉽지 않아 다중회귀모형에서도 $1/\hat{y}_i^2$ 을 사용한다.

가중회귀 추정치를 구하는 문제는 다음과 같이 생각할 수 있다. 종속변수가 y_i^*, 설명변수가 $1/x_i$ 인 회귀모형의 OLS 구하는 문제와 동일하다.

$$\min_{\alpha,\beta}\sum\frac{1}{x_i^2}(y_i-\alpha-\beta x_i)^2=\min_{\alpha,\beta}\sum\left(\frac{y_i}{x_i}-\frac{\alpha}{x_i}-\beta\right)^2=\min_{\alpha,\beta}\sum\left(y_i^*-\frac{1}{x_i}\alpha-\beta\right)^2$$

가중치를 $1/\hat{y}_i^2$ 사용했을 때는 다음 정규방정식에 의해 추정치를 구할 수 있다. 이를 가중회귀추정치이다.

$$\alpha\sum w_i+\beta\sum w_i x_i=\sum w_i y_i$$
$$\alpha\sum w_i x_i+\beta\sum w_i x_i^2=\sum w_i x_i y_i$$